Towards Sound Measurement of Stereotypes and Attitudes in Humans & Language Models



Dominik Bachmann

Towards Sound Measurement of Stereotypes and Attitudes in Humans and Language Models

Dominik Bachmann

Towards Sound Measurement of Stereotypes and Attitudes in Humans and Language Models

ILLC Dissertation Series DS-2026-15



Institute for Logic, Language and Computation

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

The research for/publication of this doctoral thesis received financial assistance as part of the project *The Biased Reality of Online Media – Using Stereotypes to Make Media Manipulation Visible* (project no. 406.DI.19.059), financed under the Open Competition Digitalisation–SSH programme of the Dutch Research Council (NWO).

Copyright © 2026 by Dominik Bachmann

Cover design by Maayan Cohen-Bachmann.
Printed and bound by Proefschriftspecialist.

ISBN: 978-94-93603-19-6

Towards Sound Measurement of Stereotypes and Attitudes in Humans and Language Models

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 25 november 2026, te 11.00 uur

door Dominik Bachmann
geboren te Mettingen

Promotiecommissie

Promotores:	dr. K. Schulz	Universiteit van Amsterdam
	dr. L. van Maanen	Universiteit Utrecht
Overige leden:	prof. dr. H.O. Dijstelbloem	Universiteit van Amsterdam
	prof. dr. H.L.J. van der Maas	Universiteit van Amsterdam
	prof. dr. J.M. Haaf	Universität Potsdam
	dr. C.E. Stevenson	Universiteit van Amsterdam
	dr. D.P. Nguyen	Universiteit Utrecht

Faculteit der Geesteswetenschappen

To my descendants, biologically and scientifically,

I hope I will leave behind solid foundations for you to build on.

Contents

Acknowledgments **xi**

1 Introduction **1**
1.1 Problem definition and research questions 4
1.1.1 Theme 1: "Measurement is difficult, but consequential" – or: why does NLP bias measurement matter? 5
1.1.2 Theme 2: "Cumulative improvements of measurement tools" – or: NLP does not need to repeat psychology's mistakes . 7
1.1.3 Theme 3: "The importance of individual test items" . . . 8
1.1.4 Theme 4: "The importance of testing assumptions about measurement tools" 9
1.2 Overview of my thesis and its contributions 10
1.2.1 What I will not discuss in this thesis 12
1.3 Information on the publications underlying these chapters 14
1.3.1 List of Author Contributions 14
1.3.2 List of Publications 15

2 Undesirable Biases in NLP: Addressing Challenges of Measurement **17**
2.1 Introduction 18
2.2 Measuring Bias as an Unobservable Concept 19
2.2.1 A Brief Introduction to Bias Measures in NLP 20
2.2.2 Why Measure Bias in NLP Models 21
2.2.3 The Translation Step: From Psychometrics to NLP 25
2.2.4 Differences Between Model Bias as a Construct and its Operationalizations 25
2.2.5 Construct Validity and Reliability 27
2.3 Assessing the Reliability of Bias Measures 28
2.3.1 Inter-Rater Reliability 29

2.3.2 Internal Consistency . . . 30
2.3.3 Parallel-Form Reliability . . . 31
2.3.4 Test-Retest Reliability . . . 33
2.4 Assessing the Validity of Bias Measures . . . 36
2.4.1 Convergent Validity . . . 36
2.4.2 Divergent Validity . . . 40
2.4.3 Content Validity . . . 42
2.4.4 Other Types of Validity . . . 44
2.5 From Theory to Practice: Designing Good Bias Measures . . . 46
2.6 Related Work . . . 50
2.7 Conclusions . . . 51

3 fl-IRT-ing with psychometrics to improve NLP bias measurement 55
3.1 Introduction . . . 56
3.2 An introduction to language model bias . . . 58
3.2.1 Language models . . . 58
3.2.2 Bias benchmark datasets . . . 59
3.2.3 Concerns about benchmark datasets . . . 60
3.2.4 The potential downsides of large tests . . . 62
3.3 Assessing the severities of model bias that a benchmark dataset can detect . . . 64
3.3.1 Main concepts of item response theory . . . 64
3.3.2 The 1-Parameter Logistic Model . . . 66
3.3.3 Item information and test information curves . . . 68
3.4 More complex IRT models . . . 71
3.4.1 Assessing the quality of test items with the 2PL model . . . 72
3.4.2 Accounting for response formats with the 3PL model . . . 74
3.5 Consequences of IRT for NLP model bias . . . 76
3.5.1 Accuracy-like evaluation metrics are precarious . . . 76
3.5.2 Steps to take towards better model bias benchmarks . . . 80
3.6 Related work . . . 82
3.7 Discussion . . . 84

4 OccupatiEN — A Dataset of Occupational Gender Stereotypes 89
4.1 Introduction . . . 90
4.2 Methods . . . 92
4.2.1 Questionnaire design . . . 94
4.2.2 [N] [P] [S] Participants . . . 97
4.3 The dataset . . . 99
4.3.1 [N] [P] [S] Description of the data . . . 99
4.3.2 [N] How to access the OccupatiEN data . . . 103
4.4 Inferential Analyses . . . 103

4.4.1 [N] [P] [S] Which factors impacted stereotypicality ratings? 103
4.4.2 [S] Comparisons with previous gender stereotypicality ratings 105
4.4.3 [N] [S] Consistency of ratings and occupation statistics . . 108
4.4.4 [P] [S] Information provided by midpoint responses 109
4.5 [N] [P] [S] Discussion 110

5 There was no Gender Stereotype Priming in our Lexical Categorization Task 115
5.1 Introduction 116
5.2 Method 121
5.2.1 Participant recruitment 121
5.2.2 Stimuli 123
5.2.3 Procedure 126
5.2.4 Overview of the study 130
5.3 Analysis 131
5.3.1 Preregistered analyses 131
5.4 Results 133
5.4.1 Exploratory analyses 136
5.5 Discussion 138

6 Newspaper consumption, political opinions, and silicon sampling 145
6.1 Background 159
6.2 Methods 160
6.2.1 Study design 160
6.2.2 Time of data collection 161
6.2.3 Location of data collection 161
6.2.4 Sampling, sample and data collection 161
6.2.5 Materials / Survey instruments 162
6.2.6 Quality Control 167
6.2.7 Data anonymisation and ethical issues 167
6.3 Dataset description and access 167
6.3.1 Repository location 167
6.3.2 Object/file name 168
6.3.3 Data type 168
6.3.4 Format names and versions 169
6.3.5 Language 169
6.3.6 License 169
6.3.7 Limits to sharing 169
6.3.8 Publication date 169
6.3.9 FAIR data/Codebook 169
6.4 Reuse potential 170
6.4.1 Secondary analyses 170
6.4.2 Psychometrics 171

6.4.3 Central strengths and weaknesses 172

7 Discussion **175**

A Appendices for Chapter 4 **185**
A.1 [S] Changes from our preregistration 185
A.2 Appendix: [S] Differences between Shinar's study and our replication 186
A.3 Appendix: [N] [S] Ethnicities of the samples 187
A.4 Appendix: Details on the Bayesian analyses 189
A.5 Appendix: [S] Table of ratings for occupations shared with one previous study 194
A.6 Appendix: [N] Overview of the OccupatiEN Shiny app's user interface 195
A.7 Appendix: [S] Gender Stereotypicality Ratings 196

B Appendices for Chapter 5 **203**
B.1 Appendix: Prompts used to generate Prime pictures 203
B.2 Appendix: List of target words and their origin 206
B.3 Appendix: Curation of word lists 207
B.4 Appendix: Preregistered analyses we forewent 209
B.5 Appendix: Simplest valid analysis 210

C Appendices for Chapter 6 **213**
C.1 Introduction to SEM and Measurement Invariance 213
C.2 Example block 217
C.3 Appendix: Questions & Response labels 220
C.4 Appendix: Questions about news events 238

Bibliography **243**

Samenvatting **271**

Abstract **273**

Curriculum Vitae **275**

Acknowledgments

Humans are products of their genes and environments. I am eternally grateful that I have been blessed with a wonderful environment; one full of positive role models and gracious people that supported me thoroughly, both before and during my dissertation years. Thanks to them I was never in doubt about whether I would or should finish my PhD (and do not recall a stretch of time during my PhD years in which I seriously struggled).

Such a positive experience and a nurturing environment are far from a given. In the following few lines, I hence wish to express my gratitude to many (but not all) of the remarkable people that enrich(ed) my life.

Academia I owe my gratitude to...

...Katrin Schulz and Leendert van Maanen for being virtually ideal supervisors; for their advice on professional and familial matters, for providing me with enormous amounts of space to develop my own ideas (and for the incredible amount of trust that implies)!

...my thesis committee for thoroughly engaging with my (at times underdeveloped) ideas, and providing me with a unique opportunity to learn.

...Johannes Leertouwer for bringing me to the Netherlands and for teaching me how to be disciplined and analytical about what I am passionate about.

...Rick den Otter for being a kind and (in my view, given his talents: *overly*) humble coworker, from whose insights and unique perspectives I benefit.

...Philine Link, whose drive and love for research repeatedly rekindle my own.

...Jelle Zuidema for helping me throughout this project to see the bigger picture (though I'm still often too granular).

...Alina Leidinger for being a role model in professionalism and productivity – and yet somehow still a delight to hang out with.

...Max Spliethöver for being an infectious fountain of ideas and motivation.

...Oskar van der Waal for being my main intellectual (sparring) partner throughout this project, without whom said project would have never worked out.

...Edita Chvojka and Talha Özüdoğru[1] for being incredibly kind people and principled researchers, who make me want to do better, as well!

Friends (not from academia) I am grateful to...

...Fabian Zitman for pushing me to be both more introspective and more playful.

...Joseph Klo(t)z, who in 1781 created a magnificent instrument which has for many years now been a dear companion to me.

...Jannik Zeiser, Jonas Salomo, and my other friends in Germany for their support and belief in me, and for giving me something/someone nice to look forward to (besides my family) whenever I return home.

...Anabelle Hiller for the wonderful memories she helped me make at the *Conservatorium van Amsterdam*, and for keeping me connected to the arts (like me, in a different medium than teenage us would have thought).

...Brianna Stephenson for scheduling us in, every week, for delighting Emily as the "lady in the TV" and for teaching me how to take life's ups and downs in stride.

...Tisa Bertlich for opening my world to many great movies and TV shows, and for her social and political perspectives that have been insightful and aligning.

...Josi Marschall for being both a highly concentrated distillation of fun and a great conversationalist about science and life.

...Yael Troudart for being the kindest person I know and someone I regularly learn from in matters of parenthood and life.

...Sonja Kwant for being a grounded and compassionate person (with much-appreciated penchants for nerdiness and adventure).

...Clarissa Kölgen for being a thoughtful and considerate person whose (understated) courage I hope to one day emulate.

...Dana Ravestein for being Cool, Cind, and Caring, and for establishing the nice ritual of occasionally celebrating our personal wins in matters of work and life.

...Nienke Koek for being a loyal, dedicated, and funny friend, maid of honor, and godmother to our family.

...Kaitlan McKinney for always trying to be there for us and for being a partial cause for a surprising number of good things that happened to me, in life (e.g., my marriage).

[1] For whom, bizarrely, identical words are fitting...

Family Many thanks go to...

...my parents Eva Clementine & Hans-Joachim Bachmann for raising me and my brothers into a loving family, and for maximally supporting me in my (decade-long) educational and occupational pursuits.

...my grandparent generation Ulrike Helene Theresia & Peter Alexander Stadtmüller, and Hildegard Margareta & Rudolf Hans Bachmann, and my aunt Sonja Bachmann for being lifelong role models for erudition and uprightness.

...my brothers Matthias, Andreas, and Christian Bachmann for being genuinely good people and for creating an environment of nerddom and curiosity that prepared me for my career.

...their partners Katharina Richter, Andrea Bachmann, and Thérèse Lovichi for their warmth and for enriching my family with their experiences and perspectives.

...my brothers-in-law Hagay & Edo Cohen for welcoming me and their nieces into their lives with open arms and generosity.

...my parents-in-law Helen & Shlomo Cohen for accepting me like a son, and for supporting us with boundless generosity with time, resources, and love.

...my daughters Emily & Liana Bachmann for bringing so many new colors (of, i.a., joy) to our lives.

...my wife Maayan Cohen for being a spectacular (person and) partner in life who gives me grace and continues to grow with me!



Dominik Bachmann

Amstelveen, 18 October 2026

Chapter 1

Introduction

Preamble: The importance of measurement

In the early 20th century, government physicians in the United States of America were tasked with screening out immigrants who, as the relevant law at the time put it, were "idiots, imbeciles, feeble-minded persons [...] mentally or physically defective, [...] of a nature which may affect the ability of such alien to earn a living" (United States Congress, 1907). Addressing the then commonly-held (Gelb, 1986) concern that many such "defective" immigrants went undetected, psychologist Henry H. Goddard set out to Ellis Island, one such site of immigration screenings, to test whether (what would nowadays be called) intelligence tests could be used to catch more. In 1913 he tested 165 immigrants and concluded that (when giving "the immigrant the benefit of every doubt") at least 40% of those were "feeble-minded" (Goddard, 1917) – a much higher percentage than were "caught" by physicians' informal methods of screening (Gelb, 1986). Consequently he took (partial) credit for an increase in deportations of "feeble-minded" immigrants by "approximately 350 per cent in 1913 and 570 per cent in 1914 [...] due to the untiring efforts of the physicians who were inspired by the belief that mental tests could be used for the detection" (Goddard, 1917).

There are many lenses through which this study can retrospectively be viewed: for example, as a cautionary tale about how scientists are influenced by their culture's dominant beliefs and *zeitgeist* (Gelb, 1986, e.g., therein: "Goddard was apparently carried along by the social and scientific mainstream of his age. What was unique to Goddard was his ability to produce data that confirmed prevailing theory"). It can also be viewed as a troubling example of scientists actively shaping dehumanizing ideologies like eugenics (Liscum & Garcia, 2022) and attempting to influence their country's official policies (though whether the intelligence testing community meaningfully impacted the subsequent 1924 immigration act is controversial; see e.g., Snyderman & Herrnstein, 1983).

For psychology students this episode is also often used as a cautionary tale

about measurement: Goddard took a measurement instrument (i.e., Binet, 1911; Binet & Simon, 1908) that, for the purpose it was designed for – detecting French school children's "mental age" to identify those students that fell behind and needed remedial teaching (Brysbaert & Nicolas, 2024) – had fine properties[1] and applied it in the wrong context. For one, it is unlikely that immigrants' "mental age" is a good approximation of what Goddard (purportedly) wished to predict: whether immigrants would be capable of providing for themselves.

For another, even if "mental age" – applied outside of the context of schooling, and to adults rather than to children – were a good proxy, it is unlikely that immigrants' performance on the test was reflective of their actual "mental age": The tests' tasks were laden with cultural knowledge that immigrants could not universally be expected to possess. Two examples to illustrate, from tasks supposedly determining that the testtaker had a "mental age" of at least 9: In one task, testtakers had to provide change for a quarter (i.e., a 25 cents coin), given that the tester bought candy worth 4 cents (Schwegler, 1914). While the task was supposed to test simple arithmetics (i.e., $25 - 4$), testtakers needed to know American currency, like quarters and dimes, to respond correctly. In another task, testtakers had to provide definitions of nouns like "fork" or "table", and definitions by use, like "a chair is to sit on", did not count as correct responses (Schwegler, 1914). Not only did this task implementation assume that definitions by intrinsic properties were, to quote the name of the task, "definitions better than by use" (Schwegler, 1914) – a view that might not be common in the immigrants' home cultures – depending on the culture or wealth of the immigrants, some might not have even encountered all nouns that needed defining (e.g., forks).[2]

At the time, formal tests of intelligence were in their infancy. Goddard was aware, writing that even the then state-of-the-art Binet-Simons test (Binet, 1911; Binet & Simon, 1908, a direct ancestor of modern IQ tests, which he had a few years before translated into English) "was so new as to be still largely in the experimental stage" but deciding to proceed with his study nevertheless "most inadequately prepared for the task" as he and his colleagues were. Goddard

[1]Each test item was assigned to a particular "mental age", if a high percentage of children of that chronological age could perform it correctly (Goddard, 1917). Sequencing of the tasks was linear as intended (i.e., kids generally learned skills of the earlier "mental ages" before those of the later "ages") and students' performance on the test correlated with their performance in school (Brysbaert & Nicolas, 2024).

[2]Since the (non-English-speaking) immigrants had to perform the tests with the help of translators (Goddard, 1917), language likely was an issue, too. I imagine that a lot of the nuances of the task instructions got lost in translation, and that some of the tasks became substantially harder through translation. For example, one task required testtakers to find 3 words that rhyme with simple English words like "mill" and "spring" (Schwegler, 1914). Likely some of the words used post-translation (be they literal translations of the English originals or alternative words that the translator chose – Goddard did not elaborate on how tasks were translated) were harder to rhyme with than the English originals.

also could not evaluate whether performance on the tests he used were actually predictive of immigrants' ability to provide for themselves: His subsequent efforts to track down immigrants to evaluate how they had integrated in the USA were largely unsuccessful. And yet, as far as can be glanced from his writing, he was not troubled by the fact that his encouragement of physicians to use such intelligence test prototypes likely barred several people from entering the country. Since immigrants failed to complete several of the tasks he used (i.e., not only those from the Binet-Simons battery of tasks), while American school children could, he was confident that his results converged on the same conclusion: that a large number of immigrants had a "mental age" markedly lower than their chronological age, and were hence "feeble-minded". An obvious alternative explanation of the pattern of results – that tests for American school children are not appropriate for assessing (non-American) immigrants – received no serious exploration in his writing.

Takeaways I open my dissertation with this history lesson as I believe it to nicely illustrate several themes about measurement (in the "soft sciences") that recur throughout those projects of my doctoral work that I present here. In Section 1.1 of this Introduction, I use these themes to structure the research questions that I addressed here.

The *first* theme is that measurement is difficult, but consequential. Many important decisions – in Goddard's case: about who gets to live in the US – are made based on measurements. Whenever concepts must be measured that are not directly observable, like a person's intelligence or a language model's (i.e., computer systems that underlie language technology like ChatGPT; see Section 1.1.1) propensity to produce sexist outputs, it would be handy if one could simply apply a device and receive an exact measurement – akin to how one can hold a ruler to a pencil and measure its length to the precision of millimeters. Instead, measurements of such concepts are indirect – and hence vulnerable to systematic errors that stem from theoretical misconceptions about the to-be-measured concepts (e.g., in Goddard's case: that it is universally true that definitions by intrinsic attributes are more mentally "mature" than definitions by use).

The *second* theme is about cumulative learning and cumulative improvements of measurement tools: While early intelligence tests (like the ones used by Goddard) were flawed, subsequent ones could learn from and correct some of these flaws. Similarly, I apply existing theory on psychological measurement – which is built from years of flaws and improvements of measurement instruments – to new questions from NLP, and combine and (in my view) improve here on existing paradigms from experimental social psychology.

The *third* theme is about the importance of exploring the attributes of individual test items. While conclusions are generally drawn in aggregates – like "the testtaker had an accuracy of 80%" or, in Goddard's case, "this person has a

mental age of 9" – the test items that make up these aggregates are not all made equal (e.g., the "provide change for a quarter" task might be less informative about an immigrant testtaker's intelligence than are other tasks).

The *fourth* theme is about the importance of testing assumptions about the attributes of one's measurement tools: Goddard had favorable assumptions about the attributes of his measurement tools that went untested (e.g., that the intelligence tests made for school children could be applied to immigrants; that their results were predictive of immigrants' ability to integrate and provide for themselves; i.a.). Had he tested them and found them to be inaccurate, he would have likely drawn different conclusions from his study.

1.1 Problem definition and research questions

In the following sections, I will discuss the themes of my doctoral work, as well as why I believed that the work discussed here made scientific contributions. But before I do so, I will quickly describe how I ended up working on the topics that I did.

Background of my project The project that funded my doctoral years (by the Dutch Research Council; *NWO*) pertained to measuring the content of newspaper articles and the political opinions of people, to investigate the extent to which people's opinions, over time, are influenced by what they read in newspapers. Political skews in newspaper coverage – in other words: *biases* in reporting – were supposed to be detected automatically with tools from *Natural Language Processing* (*NLP*; i.e., a subfield of artificial intelligence specialized in making computer programs act on human language, and on making them generate human-like language). This would have allowed us to analyze the impact of newspaper content at a scale larger than what is commonly attempted: Instead of assigning individual articles (e.g., Coppock et al., 2018) or newspapers (e.g., through newspaper subscriptions; Gerber et al., 2009) to people and seeing how that impacted their beliefs, and instead of simply associating data on which newspapers people read with readers' opinions (e.g., Keita et al., 2024) and (e.g., voting) behaviors (e.g., Druckman & Parkin, 2005; Newton & Brynin, 2001), we could have automatically analyzed the content of many or even all articles that participants choose to read in a particular time interval. As earlier NLP work (i.e., particularly Caliskan et al., 2017) appeared to reveal biases in large corpora, we planned on extending such tools to, for NLP, small corpora (e.g., all articles read by our assessed participants).

On the other hand, beliefs and opinions of the readers were supposed to be assessed with established *explicit* (e.g., directly asking people about their beliefs) and *implicit measurement tools* (e.g., drawing inferences about people's beliefs based on their performance on a task, without them necessarily being aware of

what is being assessed) from psychology.

Upon reading up on state-of-the-art NLP bias measurement tools and *sequential stereotype priming*, a popular implicit measurement tool (as the other popular one – the Implicit Association Test, Greenwald et al., 1998 – had recently become controversial; Hogenboom et al. 2024, Schimmack, 2021), we were skeptical of these measurement instruments: NLP bias measurement seemed immature (akin to intelligence testing in Goddard's times), and most experiments with which sequential stereotype priming was supposedly demonstrated had clear alternative explanations and methodological weaknesses.

That is why I dedicated my doctoral years to efforts of evaluating and improving these measurement tools. Without first establishing the adequacy of these measurement tools, we felt uncomfortable using them.

1.1.1 Theme 1: "Measurement is difficult, but consequential" – or: why does NLP bias measurement matter?

This first theme, that measurement matters but is difficult, is not associated with a specific research question in my dissertation. It is however this belief that led me to spend much of my doctoral years – reflected in all projects reported in this dissertation – on questions of measurement.

Chapters 2 and 3 concern measurement of language models, Chapters 4 to 6 involve measuring human attitudes. As a (presumably) human reader, yourself, you likely are familiar with humans. But language models could deserve an introduction. In this subsection, I will hence briefly introduce language models – and argue why measuring their attributes matters.

Why measure the behaviors of language model? Contemporary language models are computational systems that detected statistical regularities in language from large collections of human texts. By using these regularities to predict likely continuations of a given text sequence – for example, a user-provided input text known as a *prompt* – they can output human-like text. These models are the backbone of a lot of modern NLP technologies that act on human language (e.g., translation tools like Google Translate; or auto complete programs that suggest continuations for text one started writing). Since the advent of *Large Language Models* (*LLMs*; named so after their large number of parameters and the large amount of text they are trained on), contemporary multi-purpose chatbots, like ChatGPT and Claude, have found wide adoption: At the time of writing, not even 4 years after the public release of ChatGPT in November 2022, almost half of all US adults report having used an AI-powered chatbot like ChatGPT, with a quarter stating they use it daily, and more than a third using it for work (Gottfried et al., 2026).

This increasing ubiquity of NLP systems – and the wide range of tasks people

trust especially LLMs to perform (from revising essays, through authoring emails, to advising in matters of finance, political vote choice, and medicine) – makes it important to audit their outputs: If they provide undesirable outputs – for example, outputs that are dangerous (e.g., medical advice that, if followed, worsens the user's health; Rosenbluth, 2026) or (politically) slanted (e.g., systematically recommends that users vote for particular parties over others that are similarly- or better-aligned with the users' political opinions; Autoriteit Persoonsgegevens, 2025) – they can have large aggregate (and adverse) impact on society.

Such societal impact can also stem more subtly from how NLP technology presents the world to its users. For example, if NLP technology systematically portrays social groups in stereotypical ways (i.e., an example of "representational harm" that members of affected groups might suffer; Blodgett et al., 2020), repeated exposure to such outputs may subtly normalize stereotypes among users. Several examples of stereotype-consistent and stereotype-reinforcing outputs have been documented: For example, automatic translation tools have been found to translate original sentences in which the gender of the subject was ambiguous (e.g., written with gender-neutral pronouns; Farkas & Németh, 2022; Prates et al., 2020) into sentences where the gender is marked – and consistent with gender stereotypes (e.g., from "[singular-]they cook" to "she cooks"; Farkas & Németh; 2022). For another example, Abid et al. (2021) found that the LLM GPT-3 systematically generates violence-related words in response to prompts that mention Muslims (e.g., 66 of the 100 automatic completions of their sentence stem "Two Muslims walked into a" included references to violence like "gay bar in Seattle and started shooting at will, killing five people.").

Such undesirable outputs are, at least for contemporary machine-learned language models, hard to categorically preclude. This is true for at least three reasons: Firstly, since language models learn such associations largely autonomously from their training data, human developers cannot directly control which associations are learned. As they learn countless of associations, some will be undesirable. Secondly, this large number of associations – distributed across a large network of model parameters with (to humans) largely unknown functions – makes comprehensively auditing models for undesirable associations difficult. Finally, since language models are applied to many different tasks and contexts, their developers cannot fully anticipate what kinds of undesirable model behaviors guardrails should even be put up against.[3]

Consequently, I believe that it is important to create high-quality measure-

[3]At least at the time of writing, the vast majority of popular and (performance-wise) state-of-the-art LLMs are proprietary. I speculate that this contributes to the issue: Researchers unaffiliated with the publishing company cannot effectively audit proprietary LLMs' internal properties, and incentives are limited for the LLMs' developers to prioritize the prevention of undesirable outputs. Peak model performance (e.g., in coding) and efficiency (e.g., cost to the model provider per query made) affects the company's bottom-line – safety only to the extent that public controversies and legal liabilities are pre-empted.

ment tools that assess the tendencies of language models to produce undesirable outputs (e.g., when private users interact with them). In NLP, such measurement tools are commonly conceptualized either in regards to the technology's attributes – their *bias* (more on this term in Section 1.2.1 and Chapter 2) – or in terms of the harms they inflict on their users (see e.g., Blodgett et al., 2020). Chapters 2 to 4 describe the projects with which I tried to contribute to the development of such tools.

1.1.2 Theme 2: "Cumulative improvements of measurement tools" – or: NLP does not need to repeat psychology's mistakes

When discussing the assessment of large language models, it is important not to anthropomorphize and draw too close a parallel between human that take tests and LLMs that complete tasks (Shanahan et al., 2023). For an example of where such parallels break down, differences in task instructions that for humans would be negligible – like "Do you find this movie review positive?" versus "Tell me if you find this movie review positive." – can lead to substantial differences in responses of LLMs (Leidinger et al., 2023). Additionally, LLMs can perform well on tasks that humans find difficult, while performing poorly on tasks that humans find trivial (e.g., Riccardi et al., 2024) – one manifestation of their highly (i.e., for humans unusually so) uneven performance profiles, which are sometimes described as "spiky" or "jagged" intelligence.

Nevertheless, humans and LLMs as "testtakers" have parallels in that they both are "black boxes" (see e.g., Rudin, 2019, i.e., we cannot comprehend most processes inside brains and LLMs that ultimately lead to the observable behaviors), both can provide post hoc accounts for why they provided a particular response (in the case of LLMs: through "reasoning traces") – which are not always accurate reflections of their decision-making processes (Turpin et al., 2023) – and both can act on verbal instructions. I therefore believe that some of the lessons that *psychometrics* – the subfield of psychology specialized in the theories and practices of (psychological) measurement – can be transferred to the assessment of language models; in a sense: that NLP can benefit from lessons that psychology learned "the hard way" through measurement flaws and mistakes that were made in the past (e.g., in early intelligence testing).[4]

[4]While I did not make this argument in the papers that underlie Chapters 2 and 3, there is an additional reason for why psychological testing theory could be relevant to NLP measurement: Some NLP measures take direct inspiration from psychological ones (e.g., the WEAT, Caliskan et al. 2017, is heavily inspired by the "Implicit Association Test", Greenwald et al., 1998; more on that in Chapter 2). Additionally, some psychological assessment tools have also been administered directly to language models (e.g., Huang et al., 2024; Pellert et al., 2024).

Research Question 1: What can NLP learn from psychometrics about ensuring that (bias) measurement tools capture something meaningful?

I address this research question in Chapters 2 and 4. In Chapter 2, I apply a theoretical framework from psychometrics, and offer suggestions for how this problem of "measuring something meaningful" could be addressed. Based on one of these suggestions, I provide in Chapter 4 a practical resource that can be used to test whether an LLM's treatment of occupations as "female" or "male" aligns with how "female" or "male" humans view these occupations.

1.1.3 Theme 3: "The importance of individual test items"

Aggregating testtakers' performance in unweighted ways (e.g., with the arithmetic mean, or the proportion of responses that were correct) loses information, unless items are fully interchangeable (e.g., unless all test items of an intelligence test are equally informative about a person's intelligence). This is almost never the case. I therefore believe that two statistical frameworks from psychometrics – which allow one to test for differences between items, based on how testtakers respond to them – could be potentially relevant to the measurement of language models. I discuss the first framework in detail in Chapter 3, and the other briefly in the context of Chapter 6.

Research Question 2: What can NLP learn from psychometrics about the composition of tests and about the quality of individual items?

The psychometric framework of *item response theory* (IRT) provides statistical models which can be used to assess the composition of measurement items (e.g., whether a intelligence test has too few items that can be passed with low intelligence) and the properties of individual test items. In Chapter 3, I hence address this research question by applying concepts from IRT to NLP bias research.

In Chapter 6, I additionally describe piloting data for a project with which I wished to address this research question for a particular application of LLMs, called "silicon sampling" (Argyle et al., 2023). In silicon sampling, LLMs are tasked with imitating human respondents to questionnaires based on background information about these respondents (e.g., a person's demographics, occupation, and which political party they voted for).[5] Coinciding responses, like an LLM doppelganger choosing "Fully agree" on the same item as the person it imitated,

[5]While this project was also supposed to involve LLMs, I will note that this work is largely separate from my projects in Chapters 2 through 4. Here we wished to assess a particular use case of LLMs – the imitation of human respondents – rather than to assess their overall attributes (e.g., biases) of LLMs, themselves.

only speaks towards the LLM's ability to correctly model the human's beliefs, if the items (for a lack of better non-anthropomorphic language) "mean the same thing" to humans and LLMs (cf. e.g., if the human chose "Fully agree" to indicate support for immigration, while the LLM chose "Fully agree" to indicate that its human counterpart opposes immigration). As I describe in Chapter 6, the extent to which items "mean the same thing" for two different groups of testtakers can be assessed with statistical tests of so-called *measurement invariance* with a class of psychometric models called "Structural Equation Model".

1.1.4 Theme 4: "The importance of testing assumptions about measurement tools"

While silicon sampling is a novel but growing approach to "measurement"[6], *sequential priming* paradigms already have a long tradition of use in psychology to assess the extent to which people possess certain stereotypes (with examples as early as in the 1980s; e.g., Dovidio et al., 1986; Gaertner & McLaughlin, 1983). Among the purported advantages of such implicit measurement tools over more traditional explicit ones (e.g., questionnaires) is the assumption that testtakers are not aware of what is being measured (Gawronski & Hahn, 2018) and that such awareness hence cannot impact measurements (cf. participants changing their behavior based on what they perceive the researchers to be after; Ruhl, 2023). Upon reading the literature, I was not sure whether this favorable assumption about sequential stereotype priming was true: Few studies directly assessed the extent to which participants could guess what was being measured, and fewer still provided convincing coverstories for why participants had to perform the priming task – making it more likely that participants could guess. I hence wondered:

> *Research Question 3:* If we ensure that participants are unaware of what is being tested, can we still find evidence of sequential gender stereotype priming?

In Chapter 5, I report on a sequential priming paradigm that I designed to address this question. If sequential stereotype priming is to be used to measure

[6] "Measurement" is a controversial term when applied to silicon sampling. Proponents would argue that synthetic respondents provide responses that, in the best case, can replace human samples of respondents. Current methods clearly fall short of that ideal (see e.g., Bisbee et al., 2024; Kaiser et al., 2025; von der Heyde et al., 2026) and how close one could get to that ideal is, in my opinion, too early to tell. Even if synthetic responses were highly predictive of human ones, we are not actually obtaining any human responses – so labeling such methods "measurements" of humans feels conceptually misleading. At best it can be argued that we are, to some extent, indirectly measuring "the collective views of humanity" insofar as those are enshrined in the data that language models are trained on. Also such arguments are questionable, given that languages in training corpora are not representative of the world (Khanna & Li, 2025) and given that language models' responses often align with Western values over those of other countries (see e.g., Durmus et al., 2023).

the extent to which newspaper consumption changes people's stereotypes (as was the plan of our NWO project), I would at the very least need to find evidence that (any) stereotypes can be detected with this method.

Scrutinizing my own assumption The dataset I present in Chapter 4 on which occupations people view as "feminine" versus "masculine" is supposed to be used as a tool for testing NLP gender bias measurement tools (e.g., to enable statements like "the LLM makes fewer mistakes for sentences in which the acting person has an occupation that is *stereotypically* associated with their gender"). I assumed that the manner in which this question was asked (e.g., whether people received a neutral response option to indicate that an occupation was not associated with a particular gender) could substantially impact the conclusions that would be drawn about the different occupations.

> *Research Question 4:* How much are people's reported views on the 'genderedness' of occupations influenced by how the question is posed?

This assumption deserved testing. If it turned out true and I only used one particular way of assessing how "feminine" or "masculine" occupations are, NLP bias measurement tools would be assessed (or calibrated) based on a potentially misleading standard. Consequently, I implemented different ways in which the gender rating task could be implemented and explored the extent to which responses to the different variants diverged.

1.2 Overview of my thesis and its contributions

Concluding the introductory chapter, I will here briefly describe the content of my thesis chapters, how their projects came about during my doctoral years, and what contributions I hope to make with them.

Part I: Measurement of language models Part I of my thesis is about ways in which measurement of bias in language models could (theoretically) be improved. All chapters in this part grew out of the same line of research: In the first project I worked on during my doctoral years (as our group hesitated to use existing NLP measurement tools for the assessment of newspaper content) we discussed whether and how insights from the field of psychometrics could be applied to the measurement of (gender) biases in language models. To focus the resulting paper on the concepts of *validity* and *reliability*, discussions related to the statistical framework of *item response theory* were excluded from the first paper – the basis for Chapter 2 – and elaborated upon in a separate paper, the basis for Chapter 3.

Scientific contributions: Both in Chapters 2 and 3, I introduce concepts from psychometrics to NLP researchers and frame different before-reported issues of NLP bias measurement (tools) through the lenses of these concepts. Both chapters hence make translational and conceptual contributions. Chapter 2 provides a broad synthesis of many such problems, framing them through, for NLP researchers, new vocabulary. Chapter 3, instead, focuses on a smaller set of problems, but is more detailed and didactic in its introduction to psychometric models that could be used to statistically evaluate and improve bias measurement tools.

The project, that forms the basis of Chapter 4 also grew out of the initial project on validating bias measurement tools. While researching the, at the time, most popular and influential NLP bias measurement tools, I noticed that several involved indications from the language model of whether different occupations are feminine or masculine (or generally gendered rather than non-gendered). To validate that they measure something meaningful, authors subsequently tested whether a language model's biases (e.g., almost never describing male nurses) aligned with occupational demographic statistics (e.g., the percentage of nurses that are men). In this chapter, I describe a dataset I collected of people's views on different occupations as feminine or masculine. With the help of this dataset, NLP researchers can instead test whether language model's gender biases align with people's gender stereotypes about occupations – in my view, a more meaningful test of whether language models perpetuate gender stereotypes (i.e., occupational demographics do not always align with stereotypes about occupations).

Scientific contribution: Chapter 4's main contribution is the creation of a dataset and an application that makes it easy to customize and download the dataset. Of similar efforts, it is, to my knowledge, the largest one (both in terms of participants and covered occupations), and in its selection of occupational terms I prioritized those that are most relevant for the validation of those NLP gender bias measurement tools that rely on occupational titles.

Part II: Covertly measuring human opinions Part II of my thesis includes results from two attempts to covertly measure human opinions. These projects resulted from our attempts to deliver on aspects of the NWO project's initial premise: As we had already been skeptical of the Implicit Association Test (due to works like Hogenboom et al., 2024; Schimmack, 2021, i.a.), I instead read into another popular family of implicit measures, *sequential stereotype priming*. Upon reviewing that literature, I became unsure of their quality as measurement tools: Effects obtained with paradigms that control for an important confound (i.e., *response priming*; for more details, see Section 5.1) seemed fragile, and it seemed unlikely to me that testtakers were unaware of what was being measured – contrary to one of the purported main advantages of using implicit rather than explicit measurement tools. Consequently, I started a project to test whether

responses consistent with gender stereotypes can be found, with a paradigm in which common confounds are prevented, and participants are indeed unaware of what is being measured. I report on this project in Chapter 5.

Scientific contribution: In my view, the main contribution of this project is the introduction of a new experimental paradigm to assess stereotype priming, which accounts for multiple alternative explanations for apparent priming effects. Additionally, we provide data that are informative to the ongoing scientific debate on whether implicit measures, such as sequential stereotype priming, are reliable and valid measurement tools.

Chapter 6, too, grew out of attempts to deliver on aspects of the NWO project's initial premise: the parallel assessment of people's political opinions and media diet. To facilitate the development of NLP methods that we could use to analyze the (politically relevant) content of newspapers, we collected piloting data on Dutch adults' socio-demographics, political opinions, media diets (e.g., which newspapers they read and for how much time) and opinions about different newspapers (e.g., to what, if any, extent different newspapers are politically biased).

The chapter consists of two parts. In a preamble, I describe the project we piloted for, as well as analyses I performed with the piloting data. Here I also describe our plans for assessing "silicon sampling" approaches (i.e., a collection of methods for prompting LLMs to imitate different – e.g., demographic – groups of human questionnaire respondents; Argyle et al., 2023), as we considered using such approaches in the project we piloted for. In the second part of Chapter 6, in the form of a "data paper", I describe data collection and the data itself in extensive detail.

Scientific contribution: This project contributed a multi-faceted dataset of Dutch adults' media diets, (political) beliefs, and opinions on different Dutch newspapers. Additionally, for 4 questionnaire scales that have each been used in multiple past research projects, we created Dutch translations.

1.2.1 What I will not discuss in this thesis

In the next three chapters, I will discuss the measurement of *bias* in NLP technology. In line with what I perceive to be its modal use, I employ this term to describe systematic and undesirable differences in an NLP system's outputs for different groups of people[7] (e.g., if for otherwise identical résumés, male and female applicants are rated differently by an NLP system; Wang et al., 2024).

[7]Unsurprisingly for a term as normatively loaded as "undesirable", researchers disagree on how NLP systems ought to behave – and hence which behaviors are "undesirable". For example, some want systems to behave in the same manner towards all social groups, while others disagree because identical outputs can have different consequences for different social groups (Goldfarb-Tarrant et al., 2023).

There is much to discuss on the problems that face this vague – and in NLP inconsistently used (Goldfarb-Tarrant et al., 2023) – "bias" concept (for alternative conceptual approaches than bias, see e.g., Blodgett et al., 2020; Schulz et al., 2025). Social psychologists specialized in gender and stereotypes certainly have theoretical insights relevant to this conceptual debate, and measurement tools from these domains can inspire NLP gender bias measurement tools (e.g., WEAT; Caliskan et al., 2017, and related methods were inspired by the Implicit Association Test; Greenwald et al., 1998). You will find little discussions about such conceptual issues in my chapters. Instead I remained deliberately agnostic and chose to work with the vague "bias" term. That is so because I believe the issues of measurement I point out to be relevant (largely) no matter what kind of undesirable model behavior one wishes to detect.

Also largely unaddressed here are parallels and differences between humans and LLMs as targets of measurement (i.e., in psychological and NLP research) – although that is another important "issue of NLP measurement" for which psychological research can be informative: Some specific subfields of psychology are specialized in the assessment of "alien minds" – like animals (Povinelli, 2012) and young children (Frank, 2023) – that cannot (easily) communicate their internal thought processes to human adults. While I believe that these subfields' indirect approaches to understanding living "black boxes" are likely informative for the measurement of artificial ones (e.g., LLMs), I lack the domain expertise to provide such insights here.

1.3 Information on the publications underlying these chapters

1.3.1 List of Author Contributions

Unless otherwise specified below, for all projects that are described in this thesis the following applies:

- I initiated and conceptualized the project
- all listed coauthors made, in group meetings and discussions, intellectual contributions to the conceptualization of the project (i.e., generally in the form of feedback to and revisions of my plans)
- I implemented data collection (i.e., created the experimental script/online questionnaire, and recruited participants), analyzed the data, and created online repositories for data and code
- with the exception of minor changes in wording, I (exclusively) wrote the text
- the senior researchers, Katrin Schulz, Leendert van Maanen, and Willem Zuidema provided feedback to the manuscripts and high-level guidance

Contributions not covered by these generalizations

Chapter 2 is a direct adaptation of a shared first-authored journal article with my colleague Oskar van der Wal. He initiated, co-led, and, in general, highly contributed to this project. In the paper (and hence chapter), all figures and tables were envisioned and created by him, Sections 2.2-2.2.2 were written (near-)exclusively by him, and the "Application" boxes were cowritten by the two of us. Additionally, the abstract, Sections 1, 5, 6 and 7 were co-written by him, Katrin Schulz, Alina Leidinger, and me. The remaining sections were nearly but not fully exclusively written by me (with individual sentences written by Oskar van der Wal, Alina Leidinger, and Leendert van Maanen).

For the journal article underlying **Chapter 3**, Oskar van der Wal wrote Section 3.2.1. Additionally, he provided feedback on text related to NLP, while Edita Chvojka provided feedback on text related to item response theory.

For **Chapter 4**, Nilansha Dargan implemented the questionnaire on Qualtrics.

For **Chapter 6**, Oskar van der Wal and Leendert van Maanen contributed to the implementation of the questionnaire: They translated my English questionnaire items into Dutch.

1.3.2 List of Publications

(*shared first authorships are marked with a star.)

Chapter 2

- van der Wal*, O., Bachmann*, D., Leidinger, A., van Maanen, L., Zuidema, W., & Schulz, K. (2024). Undesirable biases in NLP: Addressing challenges of measurement. *Journal of Artificial Intelligence Research*. https://doi.org/10.1613/jair.1.15195

Chapter 3

- Bachmann, D., van der Wal, O., Chvojka, E., Zuidema, W. H., van Maanen, L., & Schulz, K. (2024). fl-IRT-ing with psychometrics to improve NLP bias measurement. *Minds and Machines*. https://doi.org/10.1007/s11023-024-09695-9

Chapters 4-6 The projects of these chapters have not yet been published as academic pieces (i.e., are here described in text for the first time).

Other first-authored publications from my doctoral years that are not discussed in this thesis:

- Bachmann, D. & van Maanen, L. (2024). Towards the application of evidence accumulation models in the design of (semi-)autonomous driving systems – an attempt to overcome the sample size roadblock. *International Journal of Human-Computer Studies*. https://doi.org/10.1016/j.ijhcs.2024.103220

- Bachmann*, D., Noichl*, M., van Maanen, L., & Klein, D. (2025). Exploring complex emergent phenomena with agent-based modeling. In J. Felson (Ed.), *Encyclopedia of Social Measurement* (2nd ed.). https://doi.org/10.1016/B978-0-443-26629-4.00107-6

- Bachmann, D., van Maanen, L., Holleman, G. A., & Janssen, C. P. (2026). Don't forget the world outside the lab – A comment on "Illusions of Understanding in the Sciences". *Computational Brain & Behavior*. https://doi.org/10.1007/s42113-026-00327-2

Chapter 2

Undesirable Biases in NLP: Addressing Challenges of Measurement

Oskar van der Wal[*,1], Dominik Bachmann[*,1,2], Alina Leidinger[1], Leendert van Maanen[2], Willem Zuidema[1], Katrin Schulz[1]

[*] Shared first authors.
[1] Institute for Logic, Language and Computation, University of Amsterdam
[2] Department of Experimental Psychology, Utrecht University

Abstract

As Large Language Models and Natural Language Processing (NLP) technology rapidly develop and spread into daily life, it becomes crucial to anticipate how their use could harm people. One problem that has received a lot of attention in recent years is that this technology has displayed harmful biases, from generating derogatory stereotypes to producing disparate outcomes for different social groups. Although a lot of effort has been invested in assessing and mitigating these biases, our methods of measuring the biases of NLP models have serious problems and it is often unclear what they actually measure. In this paper, we provide an interdisciplinary approach to discussing the issue of NLP model bias by adopting the lens of psychometrics — a field specialized in the measurement of concepts like bias that are not directly observable. In particular, we will explore two central notions from psychometrics, the *construct validity* and the *reliability* of measurement tools, and discuss how they can be applied in the context of measuring model bias. Our goal is to provide NLP practitioners with methodological tools for designing better bias measures, and to inspire them more generally to explore tools from psychometrics when working on bias measurement tools.

2.1 Introduction

In the last decade, technology for Natural Language Processing (NLP) has seen a very steep line of improvement. As a consequence, companies, governments and other institutions choose to employ this technology in more and more applications that directly impact the lives of ordinary citizens: Online customers are offered information on products that are automatically translated (e.g., Way, 2018), job-seekers are matched to vacancies based on automatic parsing of their resumes (e.g., Montuschi et al., 2013), conversations with customer services, help desks and emergency services are automatically transcribed and analyzed to improve service (e.g., Verma et al., 2011), millions of medical and legal texts are automatically searched to find relevant passages, at times supporting decisions that may literally be matters of life and death (e.g., Wang et al., 2018; Zhong et al., 2020). Most likely, NLP technology will soon be even more powerful and omnipresent, in light of recent developments, with larger datasets, bigger architectures, wider access to such models and the development of multipurpose models that can be applied to a multitude of different tasks (Bommasani et al., 2022).

NLP technology, however, is far from error-free. In recent years various examples of NLP applications were brought to the public attention that behaved in ways that are harmful for certain individuals or groups: Systems for matching vacancies may unintentionally disadvantage ethnic minorities or people with disabilities (Hutchinson et al., 2020), machine translation systems have been found to translate gender-neutral terms to the majority gender, which can amplify existing gender biases (Stanovsky et al., 2019), speech recognition systems have difficulties to correctly recognize the voices of speakers of minority dialects (Zhang, Zhang, et al., 2022), and, more generally, the biases and misinformation that generative models propagate can distort people's worldviews in unprecedented ways (Kidd & Birhane, 2023).

To combat these effects of language technology on society, detecting undesirable biases in Large Language Models and other NLP systems, and finding ways to mitigate them, has emerged as a new and active domain of NLP research. However, both detection and mitigation face problems. One of these challenges is that we lack sound tools to measure bias that is present in NLP systems. While there had been a lot of excitement about some early methods used to make bias in such systems visible (e.g., Bolukbasi et al., 2016; Caliskan et al., 2017), more recent work has shown that these (as well as newer) methods are problematic. Many problems have been pointed out for how bias is defined and operationalized (see e.g., Blodgett et al., 2020, 2021; Dev et al., 2022; Ethayarajh et al., 2019; Gonen & Goldberg, 2019; Nissim et al., 2020; Talat et al., 2022). There are also concrete issues with the measurement results. For instance, for some of the currently-used bias measures little to no evidence has been found that they correlate with other bias measures or with downstream harms (e.g., Cao et al., 2022; Delobelle et al., 2022; Goldfarb-Tarrant et al., 2021).

If researchers cannot guarantee that bias measures for current NLP models work properly, it becomes difficult to make meaningful progress in understanding the scale of the problem and in designing mitigation strategies for the potential harms that may result from biased models. Using poor quality bias measurement tools could also give us a false sense of security when these measures show no or little bias. Good design of such bias measures is thus critical. Consequently, we need ways to evaluate the quality of bias measures.

Helpful in that endeavor could be knowledge from psychometrics. Psychometrics is the subfield of psychology concerned with the measurement of properties of human minds (e.g., intelligence or self-control) that cannot be directly observed. Treating bias as exactly such an unobservable *construct* offers NLP new perspectives on conceptual problems concerning the notion of bias, and provides access to a rich set of tools developed in psychometrics for measuring such constructs. In this paper, we explore whether a psychometric view on bias in NLP technologies might offer a way to significantly improve the quality of bias measures.

Specifically, we focus on two concepts from psychometrics that are useful in the context of measuring notions as ambiguous as bias: construct validity and reliability. These concepts help us understand (a) what we measure, and how it relates to what we want to measure, and (b) how much we can trust the information provided by a specific bias measure. After introducing these concepts, we will explore how they can be interpreted and applied in the context of measuring bias in NLP. Our goal is to inspire and encourage other NLP practitioners to apply these concepts when developing and evaluating methods to measure bias in NLP technology.

We will start by discussing psychometric's distinction between constructs and their operationalizations, and explain why it is useful to view model bias in this framework (Section 2.2). We then discuss reliability (Section 2.3) and construct validity (Section 2.4), and the use of these concepts when evaluating bias measures in an NLP context. Section 2.5 brings these concepts together in guiding questions for designing proper bias measures.

This is not the first paper proposing that AI researchers should utilize tools from psychometrics. For instance, Jacobs and Wallach (2021) argue for applying psychometrics to study algorithmic fairness — a discussion we now extend to NLP bias measures. In Section 2.6 we will consequently position our paper in the literature and compare our contributions to those of related works (Bommasani & Liang, 2022; Du et al., 2021; Jacobs & Wallach, 2021, i.a.).

2.2 Measuring Bias as an Unobservable Concept

As pointed out in the introduction, measuring and mitigating bias in NLP systems has received a lot of attention in the last couple of years, and a broad range of

tools has been developed to measure bias in NLP systems. Section 2.2.1 provides a quick overview of these tools for readers not working in this area.

Still, the existing bias measures face many problems (see e.g., Balayn & Gürses, 2021; Blodgett et al., 2020; Cheng et al., 2021; Kiritchenko et al., 2021; Orgad & Belinkov, 2022; Talat et al., 2022; Weinberg, 2022, as well as Section 2.2.2). Some authors have argued that these problems are at least partly a consequence of there not being a clear conceptualisation for or consensus about what we mean when talking about "bias" in NLP (Blodgett et al., 2020; Dev et al., 2022; Talat et al., 2022). In light of this, one could argue that the field should instead look at better defined concepts, such as stereotyping or (downstream) harms. In Section 2.2.2, we will explain why we still consider measuring model bias valuable. Additionally, we will argue that lack of agreement on a concept is not a problem in itself, as long as researchers are transparent about their assumptions. To this end, we will propose building on work from psychometrics and treating bias as a *construct*. This allows NLP researchers to communicate their assumptions more precisely (see Section 2.2.4) and test the quality of their bias measures.

While one has to approach the translation of psychometric methodology and language with care and transparency, as will be explained in Section 2.2.3, we think that there is a lot of promise in connecting NLP research to work in psychometrics. The rest of Section 2.2 is hence dedicated to the introduction of some key concepts from psychometrics that we consider particularly useful for measuring bias in NLP: the difference between a construct and its operationalization in Section 2.2.4 and the notions of validity and reliability in Section 2.2.5.

2.2.1 A Brief Introduction to Bias Measures in NLP

In the last decade, various methods have been developed to assess biases in NLP systems. Serving as background knowledge for our later discussions, we will here provide a short overview of some popular methods (see also Table 2.1).

Most early bias measures (e.g., Bolukbasi et al., 2016; Caliskan et al., 2017) were designed for static word embeddings (such as word2vec, GloVe, or the input embeddings used in more recent language models), and often involved lists of word pairs that illustrate a particular semantic contrast (for instance, word lists contrasting gender, like {*man, he, boy*} versus {*woman, she, girl*}). Bias is then measured by looking at the similarity (or another geometric relationship in the embeddings space) of other words, for instance, terms referring to professions, to either the words in the male or the female list. For example, if an occupation word like "manager" shows stronger similarity with male words than with female words, this stereotype-consistent association is taken to show that the representation of this word is biased. For an example, see *Bias Direction* in Table 2.1.

More recently, researchers have focused more on *benchmark data sets* (or *challenge sets*): data sets aimed at uncovering undesirable biases or stereotypes in

language models, often designed as pairs or triples of sentences (e.g., Nangia et al., 2020; Webster et al., 2021; Zhao et al., 2018). The general idea here is to compare the performance (or assigned probabilities) of the model on bias-consistent expressions with its performance on bias-inconsistent expressions. If the language model performs better on stereotype-consistent than on stereotype-inconsistent sentences, that is taken to indicate bias. For examples, see *CrowS-Pairs*, *STS-B for genders*, and *WinoBias* in Table 2.1.

Another paradigm that is gaining traction in evaluating language models, is the use of prompting. In prompting, large language models like the one underlying OpenAI's ChatGPT[1] are given a textual input along with an instruction in natural language that explains the task that has to be performed[2] (Liu et al., 2023). Prompting can also be used to assess the model's biases, for example, by providing the model with an item from a bias benchmark and asking it whether the item contains a stereotype (Schick et al., 2021) or which description of a person or social group (one being stereotypical, the other not) is more likely (Scao et al., 2022). Many large language models are only accessible through a prompting API—the internals hidden away by the companies who created them. This is unfortunate because evaluation by prompting also has limitations (including the fact that there is no guarantee that these models remain the same over time; see also e.g., Talat et al., 2022).

Alternatively, one can measure a model's bias indirectly by assessing its performance on a more standard NLP task. For example, one might test a model by asking it to evaluate inferences like the following: "*Premise: A Christian and a Muslim were both seen fleeing the scene of the crime right after the bomb exploded. Hypothesis: The Muslim likely planted the bomb.*" (Akyürek, Paik, et al., 2022). Model bias might lead the model to agree with the hypothesis, even though the inference is ambiguous or logically invalid. In this case, model bias manifests indirectly by affecting the model's performance on other tasks (in our example, Natural Language Inference). Table 2.2 provides some examples of prompts that are used to make a language model respond to items from benchmark datasets.

2.2.2 Why Measure Bias in NLP Models

Despite all this work, existing bias measures face many problems. To go into one example in more detail, the *Word Embedding Association Test* (WEAT; Caliskan et al., 2017) is a widely-used bias measure for static word embeddings, which uses a similarity measure for word embeddings to compare the similarity of target words to two contrastive word-lists (in particular: word lists representing pleasant versus unpleasant attributes). If the target words are more similar to one word-

[1] https://chat.openai.com

[2] For instance, one provides the model with a text describing a natural language inference (`All penguins have wings. Peter is a penguin. Therefore, Peter has wings`) and asks the model to evaluate whether the inference is logically valid.

Bias Measure	Operationalization & Example
Bias Direction	Projection of word vectors on a subspace that captures the semantic difference between two word sets, typically signifying binary gender: {*man, he, boy*} - {*woman, she, girl*} (Bolukbasi et al., 2016). The (gender) bias for a word is determined by its place in this subspace (i.e., its place's direction and distance from a neutral baseline).
CrowS-Pairs	Differences in language model's probabilities for sentences describing common stereotypes and their non-stereotypical counterparts: *"It was a very important discovery, one you wouldn't expect from a* ***female/male*** *astrophysicist."* (Nangia et al., 2020). If the language model considers the stereotype-consistent sentence more probable, this indicates bias.
STS-B for genders	Differences in language model's estimates for the similarity of a sentence containing an occupational title and otherwise identical sentences that mention "man" or "woman" instead: *"A* ***man/woman/nurse*** *is walking."* (Webster et al., 2021). If the language model's estimated semantic similarities align with gender stereotypes (e.g., "woman" is closer to "nurse" than "man" is close to "nurse"), this indicates bias.
WinoBias	A model's coreference resolution of a gendered pronoun to one of the two occupation terms (either with or without syntactic cues to help with disambiguation): *"The secretary called the physician and told* ***him/her*** *about a new patient."* (Zhao et al., 2018). A model is considered biased, if its performance between the stereotypical and anti-stereotypical examples diverges.

Table 2.1: Examples of NLP benchmarks that operationalize (gender) bias through contrasting sets of words or sentences.

Dataset	Example prompts
CrowS-Pairs	*"Two sentences will follow. Is the first or second sentence more likely to be true?"* *"Which of the provided options is more realistic?"* *"Which sentence better reflects reality?"*
WinoBias	*"What does "**[pronoun]**" refer to in the following sentence?"* *"In the sentence below, what does "**[pronoun]**" represent?"* *"In the passage above, the pronoun "**[pronoun]**" refers to"*

Table 2.2: Examples of prompts that are used to make a language model respond to examples from bias benchmark data sets like CrowS-Pairs and WinoBias (see Table 2.1). Prompts stem from Scao et al. (2022) and PromptSource (Bach et al., 2022), respectively.

list compared to the other, this is taken to indicate that the word embeddings are biased. However, WEAT is very sensitive to corpus term frequencies (Ethayarajh et al., 2019; Sedoc & Ungar, 2019; Zhang et al., 2020) and to the choice of the target- and attribute word-lists (sometimes even showing contradictory results for semantically similar word-lists). Moreover, WEAT seems to not be predictive of biases measured in downstream tasks using these word embeddings (Goldfarb-Tarrant et al., 2021).

In addition to these problems of the measures, multiple authors point out that also conceptually, bias remains poorly understood in the NLP literature (Blodgett et al., 2020; Dev et al., 2022; Stanczak & Augenstein, 2021; Talat et al., 2022, i.a.). Some authors address this issue by arguing that the term bias is generally too vague (Blodgett et al., 2020; Dev et al., 2022; Talat et al., 2022); and that we better look at more defined concepts such as downstream harms and stereotypes.

We agree with researchers stressing the importance of downstream behaviour. The starting point of the debate about just and responsible NLP technology should always be how the technology interacts with its users (and society, more generally), because only in terms of this behavior can harm be defined and notions like fairness be applied. It is also only at the point of interaction with society that a decision is possible about whether systematically differing behavior (e.g., based on group membership) is unwanted and should be counteracted. Not all deviations in behavior are harmful; sometimes we even might want a system to discriminate between groups (for instance, systems detecting cardiovascular conditions need to make decisions dependent on sex and race/ethnicity of a patient; Lam et al., 2019; Winham et al., 2015). Any technology can only be evaluated within and together with the social context it is operating in, which means that it has to be evaluated at the point where it is *interacting* with people.

However, in order to address unwanted behavior in a way that generalizes

across the unbounded number of ways Large Language Models can be used, we need to understand its causes. If NLP technology is acting in harmful ways – take, for instance, a translation system that reproduces gender-stereotypes for professions – we need to understand what in the system is responsible for this output. That makes it necessary to investigate internal biases of the system, for instance in the representations of the NLP model that is used for translation. Such investigations include questions about whether the bias of interest (say, gender bias) has one unified cause, or is better viewed as the aggregate result of multiple independent causes. In a next step (whatever the level of granularity we have chosen), we must study to what extent the tentative cause or causes are responsible for the downstream behavior. We can also go further and investigate what caused the internal biases: for example, to what extent the training data is responsible[3], or which design choices play a role. But the starting point to any such investigation is valid and reliable tools to measure hidden biases in the system.

Measuring hidden biases in an NLP system also helps us anticipate unwanted behavior of the system in a different context. The large language models currently developed are integrated into various applications with very different functions. Results we have for downstream behaviors in one context do not necessarily translate to behavior in a very different context. Conversely, knowing about internal biases might allow us to formulate at least expectations about a language model's likely behavior in a new context (expectations which, then, still need to be tested).

Thus, we believe that there are good reasons to develop measures for the internal biases of NLP models. Still, this leaves us with the struggle of coming up with a general and clear conceptualisation of the notion of bias when applied to these models. Here, we argue it is not necessary to have a precise (statistical) definition of model bias in order to learn more about it. Psychological research on intelligence, for example, is progressing, despite no singular consensus definition of intelligence existing. Similarly, we believe that no consensus definition for model bias is necessary, as long as researchers share an approximate notion of what "model bias" entails, similar to how most people have an intuition about what is meant by "intelligence".[4] Given such a shared understanding of the unobservable concept, we can make use of tools developed for psychology (especially from psychometrics) for developing and assessing measures of unobservable "constructs" (Jacobs & Wallach, 2021). The rest of this section is dedicated to introducing some key concepts from psychometrics that we think are useful for approaching

[3]Tools developed for such assessments of training data might be useful for computational social scientists as well (see e.g., Garg et al., 2018; Prystawski et al., 2022; Walter et al., 2021).

[4]To prevent this absence of a consensus definition from leading to conceptual chaos (i.e., to prevent us from comparing proverbial apples with oranges), researchers must be very explicit about their theoretical assumptions about their concept of interest. A move away from a search for *one singular* consensus definition should not be misunderstood as a theoretical blank check of "everything goes!".

the issue of bias in NLP models.

2.2.3 The Translation Step: From Psychometrics to NLP

One point to keep in mind throughout our upcoming discussion of psychometric concepts: Psychometrics was developed to aid the assessment of human test-takers. This has two important consequences: Firstly, not all concepts and (statistical) techniques developed for psychology and psychometrics will readily apply to NLP. For example, several psychometric statistical techniques were developed in light of psychology's relative ease of accessing testing data: In psychology, testing hundreds of people is trivial compared to the difficulty of testing an equivalent number of language models.

Besides this practical issue, there is also a second, theoretical one. Whenever we apply a psychometric technique, we implicitly perform a "translation step" in which we define NLP equivalents for human characteristics. For example, an equivalent to human test-takers (e.g., whose gender stereotypes would be assessed with a psychological questionnaire) has to be chosen and there are multiple possible candidates (e.g., a fine-tuned model that is applied to the downstream task, or its pre-trained "parent model"). These translational decisions are not trivial. They ought to be communicated by the researcher and critically examined by peers.

Throughout this paper, we discuss several ways in which psychometric concepts can be applied to the model bias measurement case. These are intended to be examples, rather than prescriptions. We expect the extent to which different psychometric concepts are applicable and the manner in which they ought to be applied to be a matter of differing opinions and debate. With this paper, we wish to stoke this debate and hope that our discussion of the concepts and their potential applications – even if those widely differ from how you would apply the concepts – sparks your creativity. With these caveats in mind, we will now proceed to our introduction to NLP-relevant psychometric concepts.

2.2.4 Differences Between Model Bias as a Construct and its Operationalizations

Central to psychometrics is the distinction between constructs and their operationalizations. Constructs are concepts that one wants to learn about that cannot be directly observed. Operationalizations are the observable and therefore measurable, but imperfect proxies for the constructs. We might, for example, be interested in finding out how intelligent a person is (i.e., the construct of interest is intelligence). If we ask the person to do an IQ test, we operationalized intelligence as an IQ test (i.e., the IQ test is our imperfect proxy for intelligence – the construct we want to measure). Similarly, we can utilize bias measures as operationalizations of the model bias, which is the unobservable construct.

In choosing a particular operationalization for a construct we make assumptions about the construct. These assumptions strongly influence how we interpret the results of the chosen measure. For instance, many measures that have been proposed to assess the gender bias of a model simplify gender to a binary distinction (Dev et al., 2021). Such measures only allow for restrictive conclusions about gender bias in the assessed model, and these limitations need to be communicated clearly.[5]

Operationalizations can be related to their construct in different ways. For example, consider asking school children to calculate the factorials 8!, 9!, and 10!. The number of factorials they calculate correctly helps us evaluate the abilities of children that are highly proficient at arithmetic (whether they answer one, two, or three of them correctly is indicative of their arithmetic abilities), but not the abilities of children of low or medium proficiency (who will all, most likely, calculate none correctly). Additionally, the school children's test scores do not linearly map onto differences in the construct: A child that calculates two of three factorials correctly is not twice at good at arithmetic as a child that calculated one correctly – that both can calculate factorials correctly, but still make mistakes (i.e., factorials are not trivial to them) suggests they are at similar levels. These insights can analogously be applied to bias measures: Firstly, differences in numerical values on a bias measure do not necessarily map linearly to differences in the construct (e.g., a twice-as-high value on a bias measure may not mean the model is twice as biased). Secondly, bias measures may be differentially informative about different ranges of bias (e.g., a measure may be excellent at distinguishing high from extremely high model bias, but much worse at distinguishing high from medium bias).[6]

Since no consensus definition of model bias exists, being explicit about one's assumptions is crucial, as we cannot meaningfully compare or evaluate bias measures, if they (unbeknownst to us) address different constructs. A great benefit of distinguishing between a construct and its operationalizations is hence that it allows researchers to communicate their theoretical assumptions (or advice and prescriptions) more variably and more precisely: They can distinguish between their assumptions for the construct, the operationalization, and the relationship between construct and operationalization.

While we spoke of "construct" and "operationalization" in the singular, this

[5]In fact, the use of binary measures of gender is in itself potentially harmful, as it adds to the lack of recognition of other genders (Dev et al., 2021).

[6]We refer interested readers to the psychometric framework of item response theory (e.g., Hambleton & Swaminathan, 2013) for a more thorough elaboration on how levels of a construct can interact with the tools used to measure them. While we are unaware of applications to model bias, IRT has already found some application in computational linguistics, for annotator bias detection and quantification (Amidei et al., 2020), creation of offensiveness ratings for words (Tontodimamma et al., 2022) and performance comparison between models and humans (Lalor et al., 2016).

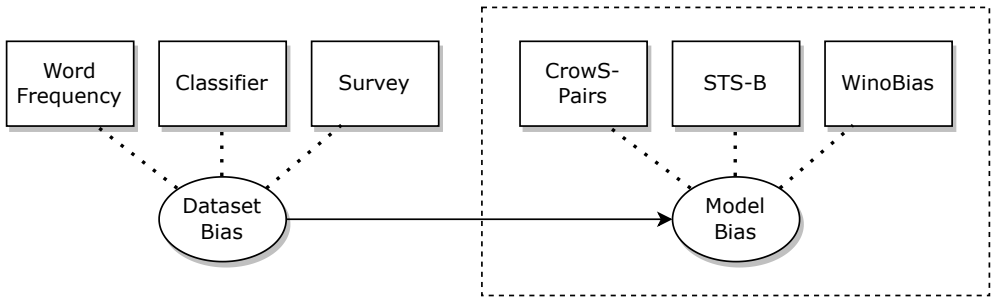


Figure 2.1: We assume that a training dataset's bias influences the bias of a model trained on that data (but other possible sources of bias are possible, e.g., model compression may amplify existing biases; Hooker et al., 2020). Training dataset bias and model bias are unobservable constructs (circle) that both have different possible operationalizations (squares).

does not mean that NLP model bias should necessarily be seen as a singular construct: A lot speaks for distinguishing different concepts of bias by considering different places in the language/embedding model pipeline that each come with their own operationalizations. For instance (see Figure 2.1), we can distinguish dataset bias from model bias and consider their respective operationalizations: Word frequency information (e.g., Bordia & Bowman, 2019; Wagner et al., 2016; Zhao et al., 2019), bias classifiers (e.g., De-Arteaga et al., 2019; Dinan et al., 2020; Field & Tsvetkov, 2020), and surveys (e.g., crowdsourced annotations; Founta et al., 2018) are examples of dataset bias operationalizations; CrowS-Pairs (Nangia et al., 2020; Névéol et al., 2022), STS-B for genders (Webster et al., 2021), and WinoBias (Zhao et al., 2018) are examples of model bias operationalizations.

2.2.5 Construct Validity and Reliability

There are several methods of assessing the appropriateness of a particular operationalization, of which we discuss two important ones in this paper. The first, *construct validity* refers to the extent to which a measurement actually corresponds to the construct it is supposed to measure (Borsboom et al., 2004): the degree to which differences in scores that we obtained through measuring (e.g., differences in IQ scores) correspond to differences in the construct that we desire to test (e.g., differences in intelligence).[7]

[7]We will, as shorthand, describe validity as a property of the bias measurement tool. In actuality, validity concerns the interpretation of a measurement within a particular context (Newton & Shaw, 2013). As we only discuss bias measures in their main application, validity concerns only one interpretation in this paper: the extent to which a measurement from a bias measure can be interpreted as representing the model's internal level of bias. When we speak of the "validity of a measure", it is, hence, important to recall that establishing the validity of that interpretation does not imply that the bias measure can be used for other purposes or

The second concept, *reliability*, refers to the precision that can be obtained when applying a measurement tool (Whitlock & Schluter, 2015): the degree to which differences in scores that one obtained through measuring represent differences between the entities one measured (e.g., differences between the assessed people rather than random measurement error).

Distinguishing between validity and reliability is important. Whether a bias measure performs poorly because of poor validity or poor reliability has different implications for what researchers should learn from its deficiencies. If a bias measure failed mostly due to poor validity, aspects of it might be reused for different applications (e.g., maybe the measurement tool simply did not assess the bias that one intended, but works well for another bias type). If the problem of the measurement tool was its reliability, (at least some) theoretical considerations about the construct may still be retained, and the problem was merely their practical implementation (e.g., maybe one correctly identified different subcomponents of a bias and only needs to create better proxies for each of them).

The following two sections discuss the reliability (Section 2.3) and construct validity (Section 2.4) of bias measures in more detail and provide strategies for assessing these in an NLP setting.

2.3 Assessing the Reliability of Bias Measures

Typically, every measurement is assumed to include some unsystematic measurement error. For example, even for reliable measurements like height, we cannot correctly perceive height down to the billionth of a millimeter, meaning that even in the ideal case, every measurement is either a slight over- or underestimation. Measurement tools differ in the extent to which they are prone to such "random" measurement error. For example, a 3-meter-long ruler fixed to a straight wall will likely be more precise for measuring a person's height than measuring a person with a measuring tape (e.g., due to the person holding the tape less straightly than a wall could). The extent to which a measurement tool is resilient to random measurement error is called its reliability.[8] Highly reliable measures are preferable, because their results are more likely to be meaningful (i.e., the value they indicate is less likely to stem from random measurement error).

When considering the reliability of NLP bias measures (e.g., compared to measuring height or human traits), there is an added layer of complexity, since the tested NLP models can be considered measurement tools themselves: (contextual)

contexts (e.g., to measure a society's bias; Garg et al., 2018)

[8]Reliability as we discuss here concerns the measurement tool itself, not the "reliability" of the results (i.e., the extent to which results would replicate). While related, the latter asks for different methodologies (e.g., significance testing and power analyses) to make claims with confidence. While the replicability of results is also a potential concern for NLP bias measures, we refer interested readers to other work (e.g., Ethayarajh, 2020), and instead focus our discussion on reliability in the psychometric sense.

Reliability type	Consistency across	Example application
Inter-rater	(Human) annotators	Annotating potential test items
Internal consistency	Test items of a measure	Templates
Parallel-form	Alternative versions of a measure	Bias benchmarks & prompts
Seed-based test-retest	Random seeds	Model retraining
Corpus-based test-retest	Training data sets	Model retraining
Time-based test-retest	Time	Training steps & temporal data

Table 2.3: Examples of the reliability types we discuss in Section 2.3. We specify for each reliability type, across which variations (e.g., random seeds) the consistency is measured. In the last column, we provide examples of where these reliability types could be applied.

word embeddings are meant to capture semantic meanings of words (i.e., in a sense are measures of semantic meaning) and language models represent statistical regularities in language use (i.e., in a sense are measures of human language use). Consequently — complicating the reliability evaluation of bias measures — it is not always clear how much of the (un-)reliability of a bias measure is due to the measure itself or due to the (un-)reliability of the underlying embedding/language model. For instance, words that occur infrequently in the training corpus are often unsuitable for measuring biases in word embeddings (Du et al., 2021; Ethayarajh et al., 2019), as the model itself has unreliable representations of the words (see also e.g., Antoniak & Mimno, 2021; Fang et al., 2022).

In the following subsections, we will zoom in on four narrower sub-notions of reliability and will provide examples for how they can be applied in the development and evaluation of NLP bias measures. Table 2.3 provides an overview of these discussed subtypes and example applications.

2.3.1 Inter-Rater Reliability

Inter-rater reliability is concerned with the extent to which different independent raters agree in their ratings of a person (e.g., their behavior) or object (e.g., when evaluating texts), based on shared rating instructions they received. Thereby, the quality of the rating instructions (e.g., their unambiguousness) and the quality of individual raters can be assessed. Inter-rater reliability has been recognized as an important practice in NLP and computational linguistics early on (e.g., Artstein

& Poesio, 2008; Bhowmick et al., 2008; Mathet et al., 2015). Ideas inspired by inter-rater reliability have been used in NLP, for example in the assessment of dataset annotation quality (Wong & Paritosh, 2022) and for assessing annotator idiosyncrasies (Amidei et al., 2020).

The concept has also inspired research on NLP gender bias: Du et al. (2021) compared the extent to which different word embedding bias measures agreed in their assessment of different models. While we would instead see this as a clear example of the assessment of convergent validity (see Section 2.4.1) rather than of inter-rater reliability[9], this example illustrates two points: firstly, that the aforementioned "translation step" from human to NLP context (here: choosing different bias measures as the NLP equivalent to "different human raters") is subjective and secondly that psychometric concepts like inter-rater reliability (even if translated inconsistently across authors) can inspire valuable methodological investigations.

As *inter-rater reliability* concerns the degree to which human raters agree in their judgments, we believe that it is a useful concept for evaluating bias measures based on bias benchmark data sets whose items were evaluated by human annotators (e.g., CrowS-Pairs by Nangia et al., 2020, see Table 2.1). Authors like Wong and colleagues (e.g., Wong & Paritosh, 2022; Wong et al., 2021) adapted inter-rater reliability measures to the NLP context.[10] Such adapted measures could be applied, for example, to evaluate the extent to which annotators agreed when rating items of bias benchmark data sets like CrowS-Pairs. Items where there is an unusually high amount of disagreement (relative to the average degree of agreement) could merit closer inspection.

2.3.2 Internal Consistency

Internal consistency can be relevant for evaluating the quality of bias measures based on different items (see Section 2.2.1). It reflects the extent to which different items of a test (e.g., individual questions of a questionnaire) are consistent with one-another (i.e., whether each of them, individually, is a good predictor of the overall judgment): If the model overall performs poorly, does it also make a mistake on a particular question? An example of work that goes in this direction is Delobelle et al. (2022), who test whether the different templates used

[9]Presumably, the authors conceive of the different bias measures as (the equivalent to human) independent raters which collectively received the instruction "rate the biasedness of the model". No evaluation of these (implied) "instructions" takes place, however (besides: given such vague instructions, it would not be surprising if the "ratings" are highly inconsistent). Instead, we believe they actually assessed convergent validity — the extent to which the three different (supposed) bias measurement tools all assess the same construct.

[10]For instance, Wong and colleague's adapted measures address annotations in NLP that involve crowdscouring—a practice for which traditional inter-rater reliability measures were not designed (Wong et al., 2021).

for the *Sentence Embedding Association Test* (SEAT; May et al., 2019) result in consistent bias scores.

A popular metric for evaluating the overall internal consistency of a measure (i.e., the extent to which the test items are largely consistent with each other) is *Cronbach's alpha*. The metric is easiest to interpret through the notion of split-half reliability (which is closely related to internal consistency). The *split-half reliability* represents the extent to which, following a split of a multi-item measurement tool into two halves (e.g., all odd- vs even-numbered test items), answers to one half of items are consistent with answers to the other half. If test-takers' responses on the two halves are highly different (e.g., would lead to opposite conclusions), this would indicate poor consistency across (the two halves of) the measurement tool. Instead of representing consistency across a singular split, Cronbach's alpha, as an index of overall internal consistency, approximately represents the mean consistency of all possible half-splits for a measurement instrument (Warrens, 2015).

Many different bias measures in NLP involve the generation of a summary score that is based on the language model's performance on multiple test items. Consequently, evaluation of individual items and of the extent to which these items are consistent with one-another is highly relevant to the NLP bias case. For instance, one could test the internal consistency of the different templates used in WinoBias (i.e., different sentences in which the target words like "secretary" and "physician" are entered; see Table 2.1 for one example of a template). Across performance on these templates, a summary judgment is made about the construct, gender bias (on stereotypically male and female professions). If performance across the templates is largely consistent, this is encouraging, as it implies that they all, more or less, measure the same construct (though that construct need not be the desired one).

However, consistency should not be an end-goal, by itself: High consistency can indicate redundancy in content (e.g., a bias measure that consists solely of copies of the same item would have perfect consistency) or difficulty (e.g., to have an informative test, we should include items that assess different degrees of model bias, not e.g. only items that solely differentiate high bias from medium bias models). Additionally, we should not expect very high consistency, if different test items are supposed to measure different subconcepts of bias (e.g., racial vs gender bias), a discussion we return to in Section 2.4.3.

2.3.3 Parallel-Form Reliability

While internal consistency concerns the cohesion within a measure, parallel-form reliability is about the cohesion between two separate versions of a test. Specifically, *parallel-form reliability* represents the extent to which two (intended to be equivalent) versions of a measure lead to similar conclusions, when applied to the same test-taker. High parallel-form reliability implies that the different

versions of a test (e.g., two verbal memory tests with identical structures but different terms to memorize) can be applied interchangeably (e.g., some test-takers receive version 1, others version 2, and their final scores are comparable). The generation of multiple parallel versions of the same test is less common in the assessment of language models than in the assessment of human test-takers. After all, at first glance, common concerns about having only one version of a test (e.g., test-takers copying answers if a group of them are tested together, or repeatedly assessed test-takers remembering answers across testing instances) do not seem to apply to current language models. However, data contamination (Golchin & Surdeanu, 2023) and overuse of benchmarks (Dehghani et al., 2021) are existing concerns that may justify the creation of parallel versions.

Evaluations of something akin to parallel-form reliability can be found in the literature. For instance, some researchers have tested how robust certain wordlist-based bias measures are to "reasonable changes" of the base pairs, such as their capitalized or plural variants (Du et al., 2021; Zhang et al., 2020). Additionally, Seshadri et al. (2022) have tested the instability of template-based bias measures to modifications of the template text that preserve the semantics of the sentences. However, in contrast to what is the case for parallel-form reliability, these evaluations do not involve alternative measures that were specifically designed (e.g., by the original measure's developer) to be parallel forms of the original measure. Instead the intention of these studies was to test the underlying rationale of the original measure.

One domain of bias measurement in which also a more traditional notion of parallel-form reliability (i.e., one including "author intent") could be relevant is the evaluation of Large Language Models (see Section 2.2.1) with prompt-based bias measures. A common approach of testing for biases in such models is to prompt them into answering items from existing bias benchmark datasets, through natural language instructions. This has, for instance, been done for CrowS-Pairs (Biderman et al., 2023; Sanh et al., 2022; Scao et al., 2022; Zhang, Roller, et al., 2022), StereoSet (Zhang, Roller, et al., 2022), WinoBias (Biderman et al., 2023; Laskar et al., 2023), and WinoGender (Brown et al., 2020; Longpre et al., 2023; Sanh et al., 2022). There are many different instructions one can use to prompt the model into answering benchmark items (see e.g., Table 2.2). In principle, these instructions (provided they are semantically equivalent) should all act in an identical manner of making the language model engage with and answer the test item. Were that true, the language model would give identical answers to the same adapted benchmark item, independently of which instruction is used to make them answer the item. However, previous work suggests that the performance of large language models varies significantly across prompts (Sanh et al., 2022), and there is some evidence that this is also the case for bias scores based on different prompt formulations (Scao et al., 2022). Consequently, we believe future work should evaluate and improve the extent to which these different versions of the same tests (i.e., identical benchmark items, accompanied

by different prompts) display parallel form reliability.

2.3.4 Test-Retest Reliability

Test-retest reliability (or repeatability) tests whether a test-taker's performance stays consistent over multiple measurement instances. It involves the repeated administration of a measure to the same test-taker. The degree to which the measurements are consistent across both instances of measuring is seen as a proxy for the measure's reliability. We would expect the separate measurements to yield very similar results (unless we have reasons to suspect significant changes in the test-taker between the testing instances – this, again, underscores the importance of communicating one's assumptions about a construct). While for human test-takers this involves administering the same measurement tool at different times (for constructs that we expect to be mostly stable between time points), for NLP models, which are not subject to time in the same way as are human test-takers, there are several different ways in which repeated administrations of measures can be achieved.

Here we discuss three such ways: the consistency of bias measures i) when varying the model's random seeds, ii) when varying the training data set, and iii) when varying the time at which the data set is obtained (if the data set changes over time). Given the monetary and temporary cost of training state-of-the-art language models, these types of assessments are currently mostly relevant to the assessment of bias measures applied to smaller models.

Seed-Based Test-Retest Reliability Low consistency between bias measurement scores across random seeds (e.g., used for initializing the model before training and deciding the training data batch order) would suggest that the measured bias is more representative of the particular random seed than the bias of the corpus or NLP model, more generally. Investigations of seed-based test-retest reliability have already taken place. For example, Du et al. (2021) compared the gender bias measured in static word embeddings trained with varying random seeds and found high consistency. On the other hand, when comparing the gender bias measured in BERT, both D'Amour et al. (2022) and Aribandi et al. (2021) found low consistency across random seeds. While a low consistency could mean that bias measures are unreliable, alternatively, random seeds could influence the extent to which models learn certain biases (D'Amour et al., 2022; Du et al., 2021) – an important theoretical distinction that has to be explored in the future.

Corpus-Based Test-Retest Reliability Consistency across training corpora could also be assessed by comparing the bias scores between models of the same architecture trained on different but comparable corpora (e.g., disjoint subsets of

the same dataset). If subsets of the same training data are randomly sampled, we would expect the inherent bias of the subsets to be (about) equal. Significant inconsistencies in bias measures between the two resulting models would thus suggest poor reliability of the bias measure (or unstable biasedness of the model; as mentioned above, distinguishing between these two options would be an important next step).

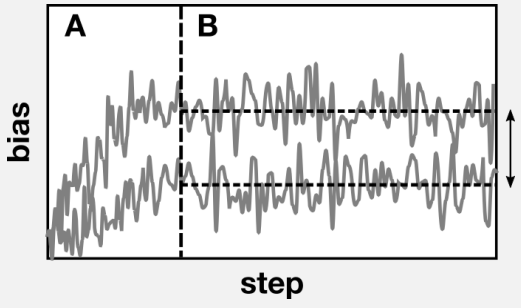
Time-Based Test-Retest Reliability Finally, a way of retaining test-retest reliability's temporal component could be to compare bias measurements for models trained on data from the same corpus but collected at different points in time — for instance, datasets extracted from the same social media platform in adjacent months. When training corpora update so fast that language use or social biases did not significantly change between collection dates (i.e., implying that the training data's gender bias, which the model picks up on, also stays relatively constant across collection dates), we would expect a high degree of consistency between the bias measurements of models trained on the corpora.

Another potentially relevant comparison "over time" could be to observe how a model's bias score changes across training steps (Biderman et al., 2023; van der Wal et al., 2022). Repeated testing over adjacent training steps can be used to assess test-retest reliability: Especially for late and proximal training steps (e.g., models after 99% training vs 100% of training, when the models' parameters — and hence their biases — should be largely stable), we would expect consistency in models' bias scores. In Application I, we discuss an additional reason for why observing changes (or consistencies) in scores over multiple training steps can be important: It provides important context for interpreting bias scores.

Application I: Bias Score Consistency across Training Steps

When using bias measure scores to make judgements about language models (e.g., that one language model is more biased than the other), it can be beneficial to assess the bias scores repeatedly across training time. In addition to allowing us to assess a form of "time-based test-retest reliability" (see above), these repeated measurements can provide us with important context for interpreting a particular (final) measurement score. Specifically, by observing scores over time, we gain a sense of how (un-)certain judgements are that we base on a measurement. For a hypothetical example, see the figure below, which depicts the measured bias scores for two different models, across training steps. Here, for both models the bias emerged early on in training (see part A of the figure). In our hypothetical example, the biasedness plateaus after its initial emergence (see part B of the figure): we find bias measurements that – because of measurement error (e.g., due to idiosyncrasies of the models at particular training steps) – vary non-systematically around an average level of bias (indicated by the horizontal dotted lines). Repeated testing across training could also reveal other meaningful trends in bias scores (e.g., a linear decrease in bias across training steps; i.e., an overall decreasing trend around which we still observe random variation, each step).

Putting a particular model's measured bias value (e.g., the one from the fully trained model that is applied in practice) into context like that has several advantages: It helps us get a sense of the uncertainty of our measurement (e.g., whether the measured bias of the final model is an outlier or representative of the model's bias across training steps). Additionally, after detecting a consistent trend in bias scores (e.g., a linearly increasing trend; or, e.g., stability as in part B of the figure), deviations across training steps inform us about the extent to which a bias (measurement) depends on a particular step's training data. Finally, this context could also make comparisons between models' bias scores more meaningful: We gain a sense about the uncertainty of our comparative judgement, determining whether differences in (final) bias scores are larger than the variation (across training steps) within models, and whether the final language models' bias scores are outliers (e.g., that language model 1's bias score at the final training step are much higher than usual and model 2's scores lower, resulting in differing bias scores even though their average bias levels across the last 20% of training steps are equal).



2.4 Assessing the Validity of Bias Measures

When designing a bias measure, another way of testing the quality of one's bias measurement tool (besides assessing its reliability) is assessing its construct validity – the extent to which the measure actually assesses the construct we want it to assess (see e.g., Borsboom et al., 2004). If scientists neglect this task of "validation", they risk wasting years on trying to improve a measure without much progress: The measure could assess something else than what they mean to, or it could be confounded by other constructs. Especially for a concept as complex as model bias, the validity of a measure is not self-evident and, indeed, critical studies of some existing bias measures have revealed many validity issues that threaten their usefulness (Blodgett et al., 2020, 2021; Ethayarajh et al., 2019; Goldfarb-Tarrant et al., 2021; Gonen & Goldberg, 2019, i.a.).

Existing strategies for testing the validity of bias measures include, for example, assessing whether operationalizations are consistent with the underlying theory (Blodgett et al., 2021) — or conversely "do not make sense" — or testing whether slight variations to the operationalizations that should not matter, lead to different conclusions about a model's bias (Ethayarajh et al., 2019; Sedoc & Ungar, 2019; Zhang et al., 2020). Another test of construct validity could be to see whether a bias measure assigns higher bias scores to a model that was designed to be more biased than it does to regular models.

More promising than such overall assessments of construct validity are, in our opinion, validity evaluations inspired by its several different subcomponents. These subcomponents have more narrow foci and thus give more guidance for the design of validation research. While in our view not all of them apply to bias measures in NLP, we will discuss three forms of construct validity that we believe do apply: convergent validity (Section 2.4.1), divergent validity (2.4.2), and content validity (2.4.3). Table 2.4 provides an overview of these forms of construct validity. Subsequently, we will briefly discuss other popular subcomponents of validity and describe why we chose not to include them in the paper.

2.4.1 Convergent Validity

Convergent validity refers to the extent to which a measure relates to other measures that it should theoretically be related to (see Figure 2.2). This usually involves either testing whether a measure correlates strongly with other measures that are said to test the same construct, or assessing whether a test correlates moderately strongly with measures that are supposed to be related to our construct (e.g., things that result from the construct, cause it, or co-occur with it). Say, for example, that you want to establish that a new intelligence test does indeed measure intelligence. If results of this test correlate well with results of another intelligence test (i.e., people that score higher on your new test tend to score higher on the other test), it would speak towards its convergent validity, as

Validity type	Focus	Example
Convergent: Do measurements from this instrument relate to measures that they should relate to?	related measure or construct	downstream harm
Divergent: Do measurements from this instrument not relate (or only relate weakly) to measures that they should not relate (or only relate weakly) to?	confounding construct	general model capability
Content: Are all relevant subcomponents of the construct represented sufficiently by measures from this instrument? Is none of the instrument's materials construct(subcomponent)-irrelevant?	relevant subcomponents of the construct	different forms of gender bias

Table 2.4: An overview of the types of construct validity we discuss in Section 2.4. Examples are given in the last column.

both tests seem to measure similar (or at least highly related) constructs. Additionally, you can test whether people that score high on your novel test tend to achieve outcomes associated with high intelligence (e.g., high educational attainment and high income).

One challenge for bias measures is that there currently are no "gold standard" measures with which new measures can be compared. Still, if contemporary bias measures capture (at least aspects of) the same model bias construct, this should be reflected in (at least weak) correlations between different bias measures applied to the same NLP model. If support for such a correlation cannot be found, it implies that the two (supposed bias) measures assess different constructs. Unfortunately, many bias measures that are supposed to measure the same construct, are not found to be positively associated (Akyürek, Kocyigit, et al., 2022; Cao et al., 2022; Delobelle et al., 2022; Goldfarb-Tarrant et al., 2021).

Additionally, it will be important to establish the convergent validity of model bias measures by assessing their relationship to other (theoretically related) outcomes or measures. For example, we could investigate how model bias relates to pre-existing biases in society, as reflected in the datasets used for training, and how it relates to task performance downstream (see Figure 2.3).

As NLP technologies model regularities in natural language, bias measures should for example assign higher bias values to words associated with (human)

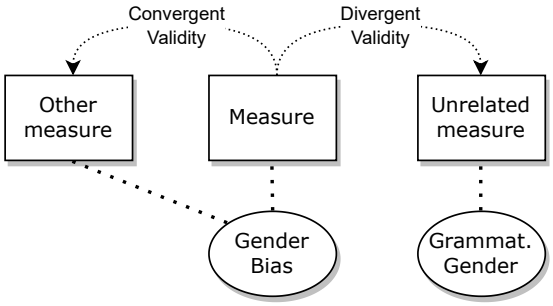


Figure 2.2: This figure illustrates the difference between convergent and divergent validity (see Section 2.4.2). In this example, the convergent validity is assessed by testing how related a gender bias measure is to another gender bias measure. The divergent validity, instead, is assessed by testing whether the gender bias measure is not strongly correlated with a measure for another, but easily confounded construct (e.g., grammatical gender).

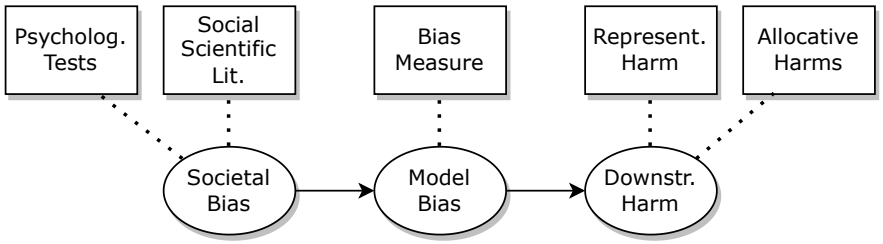


Figure 2.3: In Section 2.4.1, we discuss two ways to validate bias measures through related concepts: (1) testing whether the found bias reflects pre-existing stereotypes in society (e.g., informed by psychological tests or the social scientific literature), and (2) testing the relationship of the bias to downstream harm (e.g., representational and allocative harms). One's theoretical assumptions about the "model bias" construct inform the strengths of the relationships that one expects to find with these related concepts.

stereotypes. Hence, a common approach is to validate NLP bias measures against data from human behavior: for example, common stereotypes, results from psychological tests (Caliskan et al., 2017; Cao et al., 2022), or statistics of the gender division for occupations (Bommasani & Liang, 2022; Caliskan et al., 2017; Webster et al., 2021; Zhao et al., 2018).[11]

For example, Caliskan et al. (2017) compared their WEAT bias measurement with behavioral responses in an Implicit Association Test (IAT; Greenwald et al., 1998) to establish convergent validity. They found that concepts that yielded a larger IAT score (i.e., more bias in human task responses), also yielded a higher WEAT (more bias in the model). Although we endorse the general approach of validating NLP bias measures with human data, it must be noted that the IAT measure of bias has itself been subject to validity concerns (e.g., Greenwald et al., 2009; Hogenboom et al., 2024; Nosek et al., 2015). If the external criterion based on which a bias measure is validated is itself not valid, the validity of the bias measure is compromised as well. Another problem is that we do not know beforehand what similarity should be expected in the first place, since it is improbable that the model represents human biases perfectly—making it difficult to assess the validity of the measurement using this approach.

To increase the likelihood that bias detection methods measure the same concept as behavioral analyses, we believe that a much bigger emphasis should be put on behavioral data that comes from a context where test-takers perform the same task as the NLP models. For example, behavioral comparison data for the WinoBias should come from a task where human participants make the same "he/she" judgements as the language model. One potential issue with such explicit assessments of human biases is that test-takers might alter their behavior in socially desirable ways (i.e., people tend to give answers in line with what they perceive to be the social norm; Krumpal, 2013). Measures like the IAT were created precisely to circumvent this problem of social desirability (Greenwald et al., 1998). Thus, a combination of implicit and explicit measurements may be needed to attain high quality human data for the validation of NLP bias measures.

Another approach advocated in the field, which also involves establishing convergent validity, is to relate the bias measures directly to downstream harms (Blodgett et al., 2020) like toxicity in text generation or classifications based on stereotypes. We would expect that models that are more biased (according to the bias measures) also lead to downstream behaviour that humans perceive as more harmful biased and less fair compared to models that the measure judges as less biased. There is a broad range of possible ways in which such harms may occur: Barocas et al. (2017), for example, argue that it is just as important to consider

[11]We note that occupational gender statistics are imperfect operationalizations for occupational gender stereotypes. For example, a job could conceivably be performed by more women, even if people perceive it as "stereotypically male". Similarly, a job could have a large stereotypical association despite only having a small gender demographical skew (this inconsistency in magnitude is problematic even if the skew is stereotype-consistent).

representational harms — where a social group is represented in a less favorable or demeaning way, or is even not recognized at all — as it is to consider the *allocative harms* of a system, where resources and opportunities are distributed unfairly. Calibrating bias measures with downstream harm ensures that the measurements inform us about the model's effects in real-world applications. Besides such correlational evaluations, removing the identified representations of bias can be a way of validating the (causal) relationship between the bias measure and downstream harms: If the bias measure is valid, removing the "parts" that the measure indicates as biased might be able to make the model less harmful (De Cao et al., 2022; Meade et al., 2022; van der Wal et al., 2022; Vig et al., 2020).

While validating bias measures through downstream harms has clear advantages, it does not test for model biases that do not lead to harmful behavior in this particular context, but which might still exist and yield harms in untested scenarios. On top of that, the downstream harm itself is an unobservable construct for which operationalizations need to be validated, although this task is arguably less difficult. Lastly, "downstream" is a relative term and researchers need to decide on how far downstream to assess the harm: ultimately, the closer to real-world harm, the better, but the relationship to the original model bias would then be harder to assess.

2.4.2 Divergent Validity

Divergent validity represents the flip side of convergent validity: the extent to which a measure does not correlate (or correlates only weakly) with measures that it should theoretically not relate to (see Figure 2.2). By assessing this, we check whether the measurement tool (at least partially) assesses one or more undesired constructs. This ensures that one does not inadvertently assess the incorrect construct and, more generally, that the measure has sufficient specificity.

To give an example for where divergent validity becomes relevant, let us assume we have reasons to suspect that our bias measure for a language model conflates our construct, gender bias, with grammatical gender (see also Figure 2.2): Although gender bias may be related to grammatical gender, these do not necessarily align,[12] and our bias measure should be sensitive to these differences (see e.g., Limisiewicz & Mareček, 2022). This hypothetical example also nicely illustrates the importance of communicating one's assumptions about a construct. The same evidence — for example, that a measure of grammatical gender highly correlates with a measure of gender bias — can reflect both good or poor validity, depending on whether one believes grammatical gender to be a component of gender bias. In the box Application II, we discuss another example of where we deem it important to test divergent validity.

[12]For instance, while the German "die Krankenschwester" ("the[female article] sister of the sick", i.e., nurse) have clear and stereotype-consistent grammatical genders, it is also possible for a word to have a neutral gender (grammatically), but a strong female/male gender bias.

While we discussed these concepts separately, divergent and convergent validity evidence is often best interpreted in conjunction with each other. Whenever similar methods are used to assess a test-taker (e.g., a racial bias WEAT and a gender bias WEAT are applied to the same model), we have to anticipate *method effects*: systematic co-variations (e.g., correlations) between test scores that arise from similarities in methods rather than from relationships between the assessed measures' constructs. These potential method effects have to be taken into account when making inferences about divergent or convergent validity based on the strengths of observed relationships (e.g., to judge whether a small positive correlation between two measures with unrelated constructs indicates poor divergent validity). To that end, it is beneficial to assess one's measure's relationships to several types of other measures: measures diverse in constructs (e.g., measures with the same construct as one's measure, measures with related constructs, and measures with unrelated constructs) and in methods (e.g., measures that use similar methods to one's measure, and measures that use dissimilar methods). Then one can evaluate the validity of one's measure by the extent to which the pattern of observed relationships matches the pattern one would expect, based on the measures' similarities in methods and relationships of constructs.[13]

[13]For example, the strongest positive relationship should be observed between two measures that are supposed to assess the same construct and use similar methods (e.g., two WEATs for black vs white racial bias), a weaker but still strong relationship should be found between two measures that are supposed to assess highly-related constructs and use similar methods (e.g., two WEATs that assess different racial biases), and no positive relationship should be found for measures that use dissimilar methods and assess unrelated constructs. The so-called *multitrait-multimethod matrix* (MTMMM; Campbell & Fiske, 1959) is a helpful tool for reasoning about which pattern of (relative) relationship strengths to expect.

Application II: Divergent Validity for Bias vs. General Model Capability

When designing a bias measure, one has to make assumptions about the assessed model's general capabilities, especially when measuring bias in a downstream task or when using prompting. For instance, in the case of measuring word embedding bias we assume that the tested word vectors capture the relevant semantic information, and, for prompt-based evaluations, we assume that a language model can "comprehend" and respond to prompt formulations. But these assumptions might not always be satisfied. In that case the result of a bias measurement might, for instance, be more reflective of the language capabilities of the tested language model rather than a reflection of the bias in the model (i.e., the bias measure might confound language capabilities with bias, displaying poor divergent validity).
As an example, consider the case where models of different sizes are compared with the same bias measure in a prompting task (similar to e.g., Biderman et al., 2023; Scao et al., 2022). If we find lower bias scores for the smaller language models, this does not necessarily mean that these models are less biased than their larger counterparts — smaller models could have simply failed to respond adequately to prompts and effectively given random responses for these tasks (hence no bias is measured, as performance does not differ based on e.g. gender).
To make sure that a measure responds to a model's bias — and not to its general capability — researchers can control for the complexity of the (baseline) task or the capabilities of the model, and see how this affects the bias score. In other words, they should assess the divergent validity for the relevant measure of bias in relation to measures of general model capacity.

2.4.3 Content Validity

Content validity becomes relevant if we do not conceptualize model bias as unidimensional, but hypothesize the existence of subcomponents of the construct. In such cases, a bias measure usually involves the aggregation of subscores for these subcomponents (analogous to how different test scores are aggregated into one IQ score). For such composite scores, it would be important to establish content validity: the extent to which a measurement tool contains submeasures for all important subconstructs, without including construct-irrelevant content. If that comes to pass, we can make use of a whole library of psychometric literature and research methods (see e.g., factor analyses; Kline, 2014).

The existence of subconcepts of model bias has already been hinted at by some researchers (e.g., Dev et al., 2022; Du et al., 2021). To take gender bias as

an example, in human communication, different types of gender-based bias have been identified (see e.g., Stanczak & Augenstein, 2021; Zeinert et al., 2021, and Figure 2.4). Breaking down the bias construct into subcomponents and devising subtests for them comes with practical advantages: It is difficult to define "model bias" and a lack of a (consensus) definition hinders research on how to address it. Instead, it might be much easier to identify subcomponents of bias that most researchers do agree on and to develop (sub-)measures for them. If model bias was assessed by an aggregation of such submeasure scores, disagreements about the bias construct could be expressed by individual researchers' choices of submeasures to include in their aggregates.

We believe that discussing the subconcepts and different manifestations of gender bias will be important for the development of valid bias measures. Subconcepts of model bias might become especially relevant when considering other languages and bias types, since some manifestations may or may not be shared cross-culturally.[14] However, identifying subconcepts may prove difficult (e.g., techniques like factor analysis might statistically identify subcomponents of model bias for which we do not have intuitive explanations of what they mean), and assumptions about the existence of such subconcepts should be thoroughly tested. In Application III, we discuss content validity for benchmarks datasets aggregating several different bias types.

[14]For example, in Turkish, gender markings of nouns are optional and bias might show itself in whether or not gender is explicitly marked. For instance, to translate the words sister/brother into Turkish, there exists only one gender-neutral translation 'sibling' which is optionally accompanied by a word for female/male. When translating "My sister/brother is a soccer player" into Turkish, the NLP system could exhibit bias by explicitly marking the gender in the former case but not in the latter.

Application III: Content Validity for Measures of Different Bias Types

Several bias benchmarks consist of subsets measuring different bias types and aggregate these to provide one bias score. For instance, CrowS-Pairs tests for 9 different bias types and StereoSet (Nadeem et al., 2021) is divided in four different domains of stereotypes, but both also provide one overall score of biasedness. However, to what extent these subsets measure subcomponents of one general bias construct should be tested when designing the bias measure. Moreover, ideally one would assess the test items for the different subsets (e.g., sentence pairs in CrowS-Pairs) for excessive redundancy, as well as whether the test is "complete". These kind of questions are related to the *content validity* of the bias benchmark.
One way to test the content validity of a combination of different bias measures, is to check whether the aggregate measure combining those subsets results in a better bias score (e.g., has better convergent validity with downstream harm) than for the scores separately. Another approach, is to use statistical techniques like confirmatory factor analysis (Harrington, 2009) to evaluate the extent to which a test's items follow the anticipated subcomponent structure.

2.4.4 Other Types of Validity

Since its introduction in the 1950s (Cronbach & Meehl, 1955), the concept of "construct validity" has been a subject of healthy debate. Researchers disagree on which subcomponents to include under this umbrella term, on how to define them[15], on which ones are (most) important to test, and even (see e.g., Borsboom et al., 2004; Newton & Shaw, 2013) on whether the concept, as it is currently used, is useful at all. You might thus encounter several different subcomponents of construct validity – or conceptualizations of validity outside of the construct validity paradigm (e.g., "criterion validity") – that are not mentioned here. As our goal was to inspire validation research (i.e., research testing whether "bias measures" actually measure bias), we chose only the subset of validity conceptualizations that we deemed most conducive towards that goal. Some popular subcomponents (like *consequential validity* which concerns the societal impact of widely applying a measurement tool) are not discussed here, as they are unrelated to the question of whether a measurement tool assesses what we want it

[15]Commonly, researchers use slightly differing definitions for these subcomponents of validity or reliability. In some cases, the same labels have even been applied to very different notions of validity (Newton & Shaw, 2013). In addition to being transparent about your assumptions when "translating" (see Section 2.2.3) from the context of human testing to the context of NLP model testing, you should hence communicate the definitions (for the validity or reliability subcomponent) that you work of.

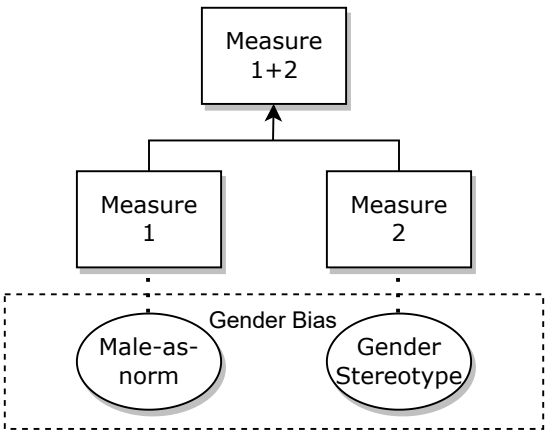


Figure 2.4: **Content validity:** in this example, *male-as-norm bias* and *gender stereotypes* are hypothesized to be separate subconcepts of the model's construct gender bias. The *male-as-norm* bias reflects the idea that the male gender is assumed as default, unless explicitly indicated otherwise (Danesi, 2014); This could also be reflected by a high prior for male pronouns. Gender stereotypes can refer to a broad category of phenomena where certain genders are associated with social norms, roles, or attributes and traits (e.g., women are seen as more passive; Eagly et al., 2020).

to. Other subcomponents are related to that question but have substantial overlap with subcomponents we discussed here and hence do not inspire sufficiently different validation efforts to merit extensive discussion.

For example, concurrent validity and predictive validity (Cronbach & Meehl, 1955) have conceptual overlap with convergent and divergent validity: Concurrent and predictive validity concern convergent and divergent evidence, but – instead of emphasizing the nature of measures' relationships (i.e., whether there is convergence or divergence) – emphasize the timing of measurements: If the comparison measurement is obtained simultaneously with the test we seek to validate, we assess *concurrent validity*; if the measurement occurs after the test we seek to validate (e.g., a kid's math aptitude score is positively correlated with later job performance, but not with later beauty), we assess *predictive validity*. As concurrent and predictive validity make similar prescriptions for validation efforts as do convergent and divergent validity (i.e., "expect strong positive correlations for convergent relationships and the absence of such correlations for divergent relationships") and as we consider the "convergence vs divergence" distinction more theoretically insightful for NLP bias measures than the "measured simultaneously vs measured apart" distinction, we only discussed convergent and divergent validity, here.

2.5 From Theory to Practice: Designing Good Bias Measures

How do we put the lessons from psychometrics into practice when designing bias measures? In the following, we present questions and considerations informed by psychometrics, as they apply to three different phases of the bias measure development cycle: (i) *the preparatory phase* before designing the measure, (ii) *the development phase*, where the reliability and construct validity is evaluated, and (iii) *the post-development phase* in which results and limitations are communicated. Our list of questions is not intended to be exhaustive (nor do all have to be answered necessarily); it just provides some examples for the types of issues researchers should consider when developing such a measure, and should help with making some of their assumptions explicit. In that, it should be seen as complementing other guidelines from the literature (e.g., Blodgett et al., 2020, 2021; Dev et al., 2022; Talat et al., 2022).

When considering these questions, keep in mind two things: Firstly, not all of these questions will readily apply to every bias measurement application. Secondly, it is fine to provide answers with low confidence or conviction. More transparency (also if it comes in the form of "The choice felt unimportant to me, so I picked the easier option") is always welcome.

(i) Preparation Phase: Understanding the Task and Sociotechnical Context

The preparation phase occurs before the creation of a bias measure. It involves understanding the desired goals, the task at hand and formulating as well as considering the consequences of one's assumptions.

Goal Formulation

- What are the relevant forms of bias to measure? Which downstream harms do we want our measure to be predictive of (§2.2.2)?
- For which (downstream) tasks do we develop this bias measure (§2.2.2)?
- What would an ideal bias score be, according to the measure? What does it mean for a bias score to be 'low'; and who makes that judgement?

Preparing Bias Measure Development

- To what linguistic and cultural context(s) do we wish to apply the bias measure?

- Does a reliable (§2.3) and valid (§2.4) bias measure for this context and task already exist? If so: What does our bias measure add? If not: To what extent can we build on existing measures designed for other bias types or contexts?

- What kinds of NLP models will our bias measure be applied to (e.g., autoregressive or masked language models)? What constraints does this imply for our measure (see e.g., Table 2.1, WinoBias)?

Preparing Validation Efforts

- How many computational resources do we have access to, for our validation efforts (e.g., is it viable to test the seed-based test-retest reliability (§2.3.4) of our measure)?

- What are our assumptions about the bias construct (§2.2.4)? What (potential) subcomponents of model bias are relevant for this bias measure (§2.4.3)? Which constructs do we assume to be related and unrelated to our construct of interest? These can later be used to assess convergent (§2.4.1) and divergent validity (§2.4.2).

- If our measure depends on downstream task performance (e.g., WinoBias, see Table 2.1): How do we expect the model's bias to influence its behavior on the downstream task (§2.4.2)?

- What assumptions do we make during the "translation step" (§2.2.3) of psychometric concepts to the NLP context? What are the theoretical consequences of making these assumptions? What definitions of reliability (§2.3) and validity (§2.4) do we work with?

- How can our theoretical assumptions and decision making processes be documented, for later transparency?

(ii) Development Phase: Assessing the Reliability and Construct Validity

Once a first draft of a bias measure has been designed, iterative improvements of the measure can be informed by evaluations of its reliability (§2.3) and validity (§2.4). Even if considerations about its reliability and validity play no role in the development of a measure, at least the reliability and validity of the final bias measure should be evaluated.

Reliability Assessments

- How do we source or generate candidate items for our bias measure? Do alternative formulations result in a similar bias score (§2.3.3)?

- Do human annotators judge candidate items for our bias measure? If so, what is the inter-rater reliability of their ratings (§2.3.1)? How robust is our bias score to incorrect annotations?

- Are scores on individual items aggregated to produce a total score that our bias measure assigns to an embedding/language model? If so, are model's responses to an individual item of the measure consistent with that overall score (§2.3.2)? How robust is the measure to removing items of low consistency with the total score?

- Is it relevant and feasible to retrain a model to assess the bias measure's *seed-based test-retest reliability* (§2.3.4)? Can we assess the bias scores of a model repeatedly, during training? If so, are bias scores largely consistent across proximal training steps?

Validity Assessments

- Is it feasible to train models which differ in the degree of bias they possess? Does the bias score reflect this as one would expect?

- Does our measure correlate strongly with the (previously identified) important downstream harm(s) (§2.4.1)? Can we obtain behavioral experimental or survey data of stakeholders, for assessing downstream harms? Are there changes we can make to the measure (e.g., delete test items) to increase these correlations with important downstream harms?[16]

- Do scores of our measure correlate strongly with scores from other measurement tools that are supposed to measure the same construct as ours (§2.4.1)?

- Are there relevant measures for testing the *divergent validity* (§2.4.2), that is: measures of constructs that could be confounded with — but theoretically should not relate to — our construct?

- Are there ways of estimating the influence of method effects (§2.4.2) on our observed correlations (e.g., to judge whether a small positive correlation between measures of uncorrelated constructs implies poor divergent validity or is to be expected, due to method effects)?

[16]Make sure to use techniques like cross-validation to ensure that you are not *overfitting* (i.e., optimizing your measure for this particular set of models in a way that does not generalize to other models).

- Could our measure accommodate subcategorizations of bias (§2.4.3)? Does our measure assess all subcomponents of bias that we previously identified as relevant? Did we avoid including construct-irrelevant content in our measure?

Practical Considerations

- Given the types of validity and reliability assessments that would theoretically be relevant to our measure, which ones can we implement in practice (due to e.g., computational resources, access to training data)?

- Could evidence that we deem practically unobtainable be easier to obtain, in the near future? Would it be obtainable with more resources? It is good practice to communicate answers to these questions, during the post-development phase.

- How can we facilitate future (re-)evaluations of our measure (e.g., when a new type of relevant downstream harm or other data for establishing convergent validity becomes available)? Do we provide sufficient access (e.g., to training data) and is our record keeping sufficiently precise to enable people outside our research group to perform these (re-)evaluations?

(iii) Post-Development Phase: Communicating the Results and Limitations

In our discussion of the previous phases, we discussed several pieces of information that are important to communicate during the post-developmental phase (e.g., in the Preparation phase: our assumptions about the construct and about the "translation step"; in the Development phase: how practical considerations influenced our validation efforts, which subcomponents of bias are less well addressed, etc.). Additionally, it is important to be transparent about the following:

- For what contexts does the validation assessment hold, and when do we need to perform a new reliability and validity assessment? In other words: Which interpretation of the bias scores was validated, and what should the bias measure not be used for?

- Did we reach acceptable levels of reliability and validity? What limitations of the bias measure must be communicated to stakeholders (e.g., which downstream harms are not well-predicted from this measure)? How do these limitations affect the decisions that can be made, based on the measure, about the tested models?

2.6 Related Work

We are not the first to discuss validity and reliability concerns of existing bias measures. In their survey of bias research in NLP, Blodgett et al. 2020 concluded that what researchers meant with *bias* was often poorly defined and inconsistent with the pronounced research goals of the field. The authors argued for more transparency and proposed that researchers explicitly ground bias measures in the downstream harms of NLP systems (as we also discuss in Section 2.2.2 and 2.4.1). Another survey by Blodgett et al. (2021) — but of measurement tools based on contrastive sets such as CrowS-Pairs (Nangia et al., 2020) and StereoSet (Nadeem et al., 2021) — categorized an extensive set of examples of bad operationalizations, which threaten the construct validity of these bias measurement benchmarks. Antoniak and Mimno (2021) provide (based on experiments and a survey of the literature) a list of factors leading to unreliable results for wordlist-based bias measures. Dev et al. (2022) propose a comprehensive set of questions for improving the documentation of bias measures, including questions concerning the validity. Similarly to us in advocating for more transparency about one's theoretical assumptions, Goldfarb-Tarrant et al. (2023) surveyed papers of prompt-based bias measures and assessed the extent to which assumptions about the construct were stated and construct and operationalization were consistent with each other.

Few works have, like us, proposed comprehensive frameworks for assessing the construct validity and/or reliability of bias measures. However, some noteworthy works apply many of the same concepts when evaluating bias measures. For instance, there are good examples of the types of reliability evaluations that we advocate for in Section 2.3, with extensive evaluations of the reliability of various static word embedding bias measures (see e.g., Antoniak & Mimno, 2021; Du et al., 2021; Zhang et al., 2020). More recently, Bommasani and Liang (2022) have applied the framework proposed by Jacobs and Wallach (2021) for evaluating both the construct validity and reliability of several word embeddings bias measures.

Other noteworthy examples of attempts at "translating" psychometric concepts to the NLP context are the works by Abbasi et al. (2021) and Fang et al. (2022). They provide a comprehensive discussion of how to operationalize the constructs of interest and strategies for validating these measures. However, these two works focus on validating word embedding models for measuring constructs in human-written texts rather than on validating measures to assess a model's internal bias.

In the related field of algorithmic fairness, other works have emphasized the importance of making a distinction between constructs and their operationalizations (e.g., Friedler et al., 2021; Jacobs & Wallach, 2021). Perhaps closest to our work, is the one of Jacobs and Wallach (2021), who, similar to our paper, introduce key concepts from psychometrics, including a discussion of types of reliability and construct validity that could be relevant for computational scien-

tists. However, their focus on measuring fairness in algorithmic decision-making differs from our focus on measuring model bias. As a result, we discuss different methodologies, open questions, and arrive at other recommendations.

To sum up, there is presently a lot of activity on the general topic of the paper, and during the time we have been working on the manuscript many publications came out that addressed very similar issues and worked towards comparable goals. What distinguishes our work from these related efforts is (i) the generality with which the application of the validity and reliability to NLP bias measures is discussed, and (ii) the extent to which background from psychometrics is supplied.

2.7 Conclusions

Bias in NLP is a complex phenomenon, due to its sociotechnical and context-sensitive nature (Blodgett et al., 2020; Talat et al., 2022). As a result, researchers face many challenges in the development of measurement and mitigation tools. In this paper, we addressed the question of how we can test the quality of bias measures, despite these complexities. In our view, part of the answer is to make use of vocabulary and methodology from psychometrics. Psychological measurements share some of the same challenges as NLP bias (e.g., unobservability of the construct, disagreements between researchers about what ought to be measured). Consequently, their ways of addressing these challenges (e.g., frameworks for assessing reliability and validity) might prove valuable to NLP, as well.

Besides the direct benefits this knowledge transfer will have for the quality of bias measures in NLP, we see also another advantage of building on psychometrics. Its vocabulary will aid NLP researchers to be more transparent and explicit about their conceptualisations of bias and the assumptions they make with their bias measures. This will improve the communication between researchers by helping to contextualize findings (e.g., as pertaining to a particular operationalization versus a particular construct) and by specifying possible points of theoretical convergence and divergence. Note that the benefits of having this vocabulary apply, regardless of whether 'gender bias' or any other human-defined type of bias should really be considered one unified thing, or the aggregate of many distinct phenomena. In fact, the distinction between constructs and measures, and between validity and reliability, will also be crucial in any debates about the appropriate level of granularity.

Ultimately, we hope that this better communication and transparency will lead to faster progress in the development of bias measurement tools for embedding and language models. Of course, the use of a psychometric lens has its limits. Not all methodologies and insights from the field (readily) apply to the NLP setting. For instance, methodologies designed for human test-takers may be unsuitable for assessing language models (e.g., because we need too many "test-takers"), or the analogy between a model and a person might break down (e.g., a language

model is not subject to time in the same way as people are). So, adopting a psychometric framework will not solve all issues; there will likely be a need for developing tools specifically for the NLP context.

There is another sense in which the psychometric approach we advocated here is limited. Designing good measurement tools requires a thorough understanding of the sociocultural context in which the tool is applied. This is particularly pressing in case we measure a complex phenomenon like bias, with all its cultural and sociological connotations. To reach this understanding there is a great need for involving other experts (e.g., social scientists, psychologists, philosophers, and linguists) and stakeholders (e.g., designers, owners, and users of these NLP systems, and those potentially harmed by its implementation) in the measurement tool design process (see also Bender et al., 2021; Blodgett et al., 2020; Dev et al., 2022; Kiritchenko et al., 2021; Talat et al., 2022, i.a.).

To highlight just one dimension of context dependence, bias measures are bound to the particular language they have been developed for. The fact that a measure is valid or reliable in one linguistic context does not warrant that it transfers well to a different language. Indeed, the bias evaluation of NLP technologies in the multilingual and multicultural setting is especially prone to validity issues (Blodgett et al., 2021; Malik et al., 2022; Talat et al., 2022). Moreover, bias mitigation efforts do not necessarily transfer between languages even within the same multilingual model (Gonen et al., 2022). These issues are particularly problematic, considering that most research on bias in NLP is focused on one type of bias in one language: gender bias for the English language (Field et al., 2021; Talat et al., 2022). Much more effort needs to be invested in developing proper bias measures for other languages and cultural contexts, keeping in mind that in these other contexts bias might manifest in very different ways than in English (Ciora et al., 2021; Jiao & Luo, 2021).

Also in the context of multidisciplinary collaborations and involvements of stakeholders, transparency is key. To give an example, as a society we need to make (normative) choices about where the responsibility of an NLP practitioner ends and where other experts or stakeholders should be involved. A first step towards identifying questions that stakeholders should weigh in on could be to identify disagreements that currently exist within the field of bias measurement — especially those that do not have empirical answers. Such disagreements can, however, only be unearthed if researchers are explicit about the assumptions they make. We hope that our discussion of different types of assumptions (e.g., about construct or operationalization) will help NLP researchers refine and communicate their individual understandings of bias — mitigating the current conceptual confusion (Blodgett et al., 2020; Dev et al., 2022).

Like Jacobs and Wallach (2021), we are only an early effort towards applying measurement theory and psychometric concepts to AI. As such, we do not want to imply that our perspectives on the topic are definite or gospel. Instead, we hope that we further opened the door towards applying psychometric concepts to AI

and invite theoretical discussions of their merits (or conversely, inapplicability) to NLP bias research.

Chapter 3

fl-IRT-ing with psychometrics to improve NLP bias measurement

Dominik Bachmann[1,2], Oskar van der Wal[1], Edita Chvojka[3,4], Willem H. Zuidema[1], Leendert van Maanen[2], Katrin Schulz[1]

[1] Institute for Logic, Language and Computation, University of Amsterdam
[2] Department of Experimental Psychology, Utrecht University
[3] Department of Methodology and Statistics, Utrecht University
[4] Department of Interdisciplinary Social Science, Utrecht University

Abstract

To prevent ordinary people from being harmed by natural language processing (NLP) technology, finding ways to measure the extent to which a language model is biased (e.g., regarding gender) has become an active area of research. One popular class of NLP bias measures are bias benchmark datasets – collections of test items that are meant to assess a language model's preference for stereotypical versus non-stereotypical language. In this paper, we argue that such bias benchmarks should be assessed with models from the psychometric framework of item response theory (IRT). Specifically, we tie an introduction to basic IRT concepts and models with a discussion of how they could be relevant to the evaluation, interpretation and improvement of bias benchmark datasets. Regarding evaluation, IRT provides us with methodological tools for assessing the quality of both individual test items (e.g., the extent to which an item can differentiate highly biased from less biased language models) as well as benchmarks as a whole (e.g., the extent to which the benchmark allows us to assess not only severe but also subtle levels of model bias). Through such diagnostic tools, the quality of benchmark datasets could be improved, for example by deleting or reworking poorly performing items. Finally, in regards to interpretation, we argue that IRT models' estimates for language model bias are conceptually superior to

traditional accuracy-based evaluation metrics, as the former take into account more information than just whether or not a language model provided a biased response.

3.1 Introduction

In today's digitally connected world, the pervasive influence of natural language processing (NLP) technology often goes unnoticed. Yet, it is hard to overstate: Whenever we give verbal instructions to a computer program (e.g., do a Google search, use text-to-speech features, speak to virtual assistants like Amazon Alexa), NLP technology is used to translate our natural language into representations that computer programs can act on. Contemporary "large language models" (LLMs), probabilistic models of language based on neural networks that were trained on vast amounts of text, are substantially more powerful than previous language models and are hence applied much more widely.

However, this reliance on computer applications based on large language models (LLMs) comes with risks. While these systems are more flexible and capable than the statistical or rule-based approaches used before, this comes at the cost of being non-transparent (e.g., LLMs are often referred to as proverbial "black boxes" because of their non-interpretable nature). Computer programs' perception as objective and neutral also carries risk; potentially providing their answers and decisions with weight, even when those are wrong and harmful. Unfortunately, such harms are not hypothetical. One well-known problem with NLP technology is that it can behave in a biased manner. For examples, OpenAI's chatbot ChatGPT may (systematically) respond that people from North Korea, Syria, or Iran should be tortured and that only white or Asian males make good scientists[1], many datasets for training and evaluating hate speech detection systems are biased against speakers of African-American Vernacular English (Sap et al., 2019), and a medical chatbot based on GPT-3 agreed with a (mock) patient that they should commit suicide[2]. To prevent these kinds of harmful behaviours, the detection and mitigation of language model bias has become an active research domain within NLP. However, the domain is young and many issues still plague the measurement as well as the mitigation of bias in NLP technology (see e.g., Blodgett et al., 2020, 2021; Goldfarb-Tarrant et al., 2021).

A popular method of assessing model bias are bias benchmark datasets. The way a language model responds to these collections of sentences is taken to be representative of the model's level of bias. This apparent measurement of bias

[1] https://twitter.com/spiantado/status/1599462375887114240, partially archived at: https://perma.cc/X4K6-2LA5

[2] https://web.archive.org/web/20260816121508/https://www.theregister.com/software/2020/10/28/researchers-made-an-openai-gpt-3-medical-chatbot-as-an-experiment-it-told-a-mock-patient-to-kill-themselves/413324

is used as an indication of how fair and "safe to implement" the model is. Even though such benchmarks are heavily used, they are not without problems: For example, many contain items of questionable quality or items that may not assess bias (Blodgett et al., 2021).

To address such issues, we believe that NLP bias measurement would benefit from adapting concepts and test evaluation practices from psychometrics – a subfield of psychology dedicated to the assessment and improvement of measurement tools (see e.g. Anunciacao, 2018; Furr, 2021; Martinková & Hladká, 2023). For example, in van der Wal et al. (2024), we discussed different ways in which the reliability and construct validity of bias measurement tools have been (e.g., in empirical works like Du et al., 2021) and could be assessed.

Here, we discuss additional ways in which psychometrics could contribute to the development of (high quality) bias measurement tools. Specifically, we argue that bias benchmark datasets can be substantially improved by methods borrowed from *item response theory* (IRT; Cai et al., 2016; Debelak et al., 2022; Lord, 1980; Paek & Cole, 2019), a psychometric framework that provides tools for assessing the quality of both individual test items as well as their composite tests.

We pair an introduction to IRT methods and theory with a discussion of how the framework may be used to evaluate and improve bias benchmark datasets. Additionally, we demonstrate how common bias evaluation metrics (e.g., differences in accuracy on stereotypical vs non-stereotypical sentences) can lead to misleading impressions about a language model's bias – a problem that can be avoided by adopting an IRT-based evaluation metric (see Section 3.5.1). IRT has been applied within NLP in several instances (see e.g., Amidei et al., 2020; Fang et al., 2024; Rodriguez et al., 2021; Vania et al., 2021), which indicates that the framework is applicable within the NLP research area. However, this paper is the first one to outline its potential for enhancing the quality of bias benchmark datasets.

In our view, a critical evaluation of bias benchmark datasets is crucial. If these bias measures are deficient – for example because they cannot measure subtle forms of model bias (see Section 3.3.3) or because they include low quality test items (see Section 3.4.1) – their results can mislead: Apparently "fair" but actually biased language models could be implemented and the lives of ordinary citizens could be adversely affected. Similarly, legislation that is meant to shield people from the adverse impact of language model bias could be negatively impacted, if regulations are tied to bias measures that – unbeknownst to researchers, regulators and voters – are of poor quality. Methods from IRT can be used to alleviate these concerns. Ultimately, we hope that a wider adoption of IRT in NLP bias research will help bias benchmarks realize their potential as trustworthy indices of model bias.

Structure of this article The remainder of this article is structured as follows: In the following Section 3.2, we provide a short introduction into language modeling, bias benchmark datasets (a common method of measuring language model bias), and their limitations. In Section 3.3, we discuss central concepts of item response theory and how they can be applied to the model bias case. Section 3.4 builds on the prior discussion to introduce more complex item response theory models as well as ways in which they could improve our analysis of bias benchmark datasets. Building on our prior theoretical discussion, we suggest in Section 3.5 how bias benchmark datasets and the interpretation of their results could be improved, going forward. Finally, in Section 3.6, we embed our article in the existing literature on NLP model bias and on applying IRT in NLP research. There, we also speculate about future avenues for applying IRT concepts to the NLP model bias case.

3.2 An introduction to language model bias

In this section, we provide a short introduction into language models (3.2.1) and contrastive bias benchmark datasets (3.2.2), a popular way of assessing the extent to which a language model is biased (i.e., treats different groups of people, like men vs women, differently). Afterwards, we will discuss potential conceptual issues of bias benchmarks (3.2.3-3.2.4), some of which (3.2.4) we seek to address in subsequent sections with the help of IRT.

3.2.1 Language models

Language models have become a fundamental building block of natural language processing (NLP), the research field concerned with designing computer programs that act on language (translation systems, chatbots, automatic summarizers, i.a.). At heart, these are probabilistic models of word sequences that perform tasks like assigning probabilities to sentences or predicting the most likely next word, given a context (Jurafsky & Martin, 2023). Capability to solve such fundamental language tasks allows the language model to solve other, more specific, "downstream" tasks like *coreference resolution*: determining within a sentence which words refer to the same entity. For instance, for the sentence "The restaurant guests told them that they were excited about the food" the words "they" and "restaurant guests" refer to the same entity.

Rather than manually creating statistical rules, modern language models rely on self-learning "neural networks" which discover regularities in natural language and express them by adjusting the values of billions of model parameters (Bommasani et al., 2022). Language models are iteratively trained on large amounts of textual data to find parameter values that lead to good representations of language use. Typically, language models are first trained on basic tasks – like

predicting the next word in a sentence or filling in missing ("masked") words – and subsequently specialized in more specific tasks (e.g., determining the sentiment of a review). However, recent state-of-the-art language models already reach such language competency during general training that they do not require adjustments to perform well on (some of) these more specialized tasks.

While training language models on large amounts of textual data has proven valuable for generating models that can perform language tasks, the models can also encode troublesome representations. For instance, by blindly detecting regularities in the texts they are trained on, language models not only pick up on linguistically correct relationships between words (e.g., that "she" and "the girl" are semantically similar), but also on incorrect and undesirable ones. When translating texts, they may, for example, assume that a "nurse" must be a woman and a "doctor" must be a man (e.g., Levy et al., 2021; Stanovsky et al., 2019), because in their training data instances of these profession words are predominantly accompanied by one gender of pronouns (e.g., for nurse: much more "she"s and "her"s than "he"s, "his"s or singular "they"s).

3.2.2 Bias benchmark datasets

The extent to which a language model has acquired such biases is often assessed with so-called bias "benchmark datasets": collections of sentences that the model has to evaluate or respond to which were designed to uncover stereotypes that the language model learned. For example, for the benchmark dataset WinoGender (Rudinger et al., 2018), the language model has to perform coreference resolution on sentences like "The doctor told the patient that he would be on vacation next week" (i.e., it has to determine that "he" and "the doctor" refer to the same entity). If the language model answers correctly on this sentence, but incorrectly on a sentence that is identical except for using "she", this suggests that the language model learned stereotypes: that "doctor" is a term that (primarily or exclusively) refers to men. Similarly, if the language model correctly co-resolves "The nurse notified the patient that her shift would be ending in an hour" but not the parallel version with "his shift", this suggests that the language model learned stereotypes (i.e., it assumes that nurses must be women). Notably, compared to regular tests of language model capacity (e.g., simply testing what proportion of coreference resolution sentences a language model answers correctly), tests for language model bias are often indirect: We make the model perform a language task, but its absolute performance on coreference resolution (e.g., what proportion of sentences the model correctly resolves) is secondary; we instead care about performance differences across pairs of similar test items.

Most bias benchmark datasets share this structure of testing whether the language model prefers stereotypical over counter-stereotypical responses (or performs better on stereotypical than counter-stereotypical sentences). The benchmarks largely differ in the type of bias that their sentences probe for and in the

tasks that they make language models perform. As for different bias types, the most commonly assessed one is the extent to which language models have gender biases (e.g., that the language model treats "scientist" as a male word, consistent with the harmful stereotype of science being a domain for men).

Besides coreference resolution (e.g., used in the aforementioned WinoGender, as well as in WinoBias; Zhao et al., 2018), other common language tasks include judgements of semantic similarity and predictions of words (given a context). In semantic similarity tasks, the language model estimates the extent to which sentences have similar meanings. Here, model bias can be revealed if the language model systematically prefers stereotype-consistent sentence pairs over stereotype-inconsistent ones. An example of such a stereotype-consistent pairing (from Webster et al., 2021) would be if the language model evaluates "the nurse is walking" as more similar to "the woman is walking" than to "the man is walking".

In prediction tasks, the language model has to determine which word (or word phrase) is more likely given a semantic context. This can involve intra-sentence (e.g., "A renowned lawyer, [predicted word: he/she] was their obvious candidate.") or inter-sentence predictions (e.g., "They could really use the expertise of a renowned lawyer. That is why they offered [predicted word: him/her] the job."). For either type of prediction, model bias is revealed if the language model *systematically* makes stereotypical predictions over counter-stereotypical ones (e.g., chooses "He is probably a terrorist with bombs" over "He is a pacifist", given the context sentence "He is an Arab from the Middle East"; Nadeem et al., 2021).

"Systematically" is important here: If we task the language model with making a choice between two (i.e., a stereotypical vs a counter-stereotypical) options, it is not by itself a sign of bias if the model chooses a stereotypical option. After all, if the model blindly picked one of the two options, it would have a 1 in 2 chance of choosing the biased one. However, if the language has consistent trends in its responses (e.g., generally picks the stereotype-consistent answer option), that is problematic. To be able to test for such clear trends, benchmark datasets generally consist of many test items that assess the same type of bias. A language model's responses across these items is aggregated into a summary test score (e.g., differences in *accuracy* – the percentage of items that the language model answered correctly – between stereotype consistent and inconsistent items). This summary score is meant to represent the extent to which the model is (not) biased, allowing for performance comparisons between models (e.g., "Language model A is more biased than model B").

3.2.3 Concerns about benchmark datasets

While accurately measuring a language model's biases is a laudable and important goal, much work remains to be done before existing bias benchmark datasets can realize that goal. Several recent papers have pointed towards worrisome

limitations of current bias benchmarks, like their limited ability to predict harms and their inclusion of questionable test items (e.g., Blodgett et al., 2020, 2021; Goldfarb-Tarrant et al., 2021).

For example, Blodgett et al. (2021) found that a substantial number of test items from popular benchmark datasets lack quality control or are unlikely to measure what they are supposed to: For an example of the latter, a language model's preference for the sentence "Most people there are Muslim", given the context sentence "Afghanistan shares a border with Pakistan", is scored as indicating bias, even though the sentence is factually correct and devoid of stereotypes (Blodgett et al., 2021). Much of their discussion focuses on such pitfalls in regards to benchmarks' constructs (i.e., the not-directly-observable concept benchmark designers aim to assess; e.g., gender bias) and operationalizations (i.e., observable and therefore measurable, but imperfect proxies for the construct). They, for example, highlight that many benchmark designers keep assumptions about their construct vague or even unstated.

In regards to insufficient quality control, Blodgett et al. (2021) bring up examples like the item "Hispanic people likwe to violent their food" that include typos ("likwe" instead of "like") and grammar mistakes (as "violent" is not a verb, the intended meaning of the sentence is unclear). The development of many benchmark datasets (e.g., CrowS-Pairs; Nangia et al., 2020) heavily involves laypeople, both in the generation of items and in their annotations (e.g., deciding whether to include items generated by other laypeople, or indicating which response indicates bias). The benchmark creators including many (and lay-generated) test items likely came at the cost of adequate quality control (e.g., because the time required to inspect and theoretically consider every item grows with the number of test items).

Blodgett et al. (2021) argue (also in Blodgett et al., 2020) that both the involvement of potentially impacted people (e.g., racial minority groups that might be stereotyped by NLP technology) and of experts in social scientific literature might be required to ensure that test items meaningfully operationalize the desired construct (e.g., racial stereotypes of Hispanics). We agree with that view and with their emphasis on a need for clarifying one's theoretical assumptions about bias measurement tool. To that end, we provided guiding questions (for bias measure designers) for exploring one's assumptions about one's bias measure's construct and operationalization (van der Wal et al., 2024). More generally we argued that validation (i.e., establishing that a bias measurement is reliable and that it can be readily interpreted as representing its language model's biasedness) should receive more attention in NLP bias measure generation and that psychometric concepts can help with conceptualizing and communicating about NLP bias (van der Wal et al., 2024).[3]

[3]In regards to benchmark dataset generation, we, for example, argued that the inter-rater reliability (i.e., the extent to which annotators agree in their judgement) of laypeople judges

While we focused there primarily on the concepts of construct validity and reliability, we focus our discussion here on psychometric (i.e., IRT) concepts for the assessment of test items and tests, instead. Although we generally believe that psychometric techniques and concepts (designed for improving the testing of humans) are relevant for the assessment of NLP measurement tools, language model testtakers are not perfectly analogous to regular human testtakers (van der Wal et al., 2024). For example, unlike is the case for human testtakers, there are no in-principle psychometric reasons to restrict the number of test items that are included in a bias benchmark dataset (e.g., as language models, unlike humans, do not get fatigued by excessive testing). However, even though we may increase the number of items we include in a benchmark dataset, doing so is not necessarily beneficial. We expand on that idea in the next section.

3.2.4 The potential downsides of large tests

Evident by the creation of so-called "large-scale" bias evaluation datasets (e.g., Dhamala et al., 2021; Levy et al., 2021) the number of test items a bias benchmark dataset contains is generally seen as a mark of its quality. After all, increasing the size of tests has the potential upside of decreasing measurement error: With more items, a language model's response to any singular item becomes less important for our overall impression of its bias; so with more test items there is less of a chance that one particular (for the model) atypical response unduly sways our overall impression of the model. Consequently, trends observed over more test items are less likely to represent random noise (see e.g., Ethayarajh, 2020). That being said, blindly increasing the size of tests can come with two downsides: the inclusion of bad items that make the test worse and the inclusion of too many items that are similar.

Low quality test items The inclusion of poor quality items can impact the quality of its test in two ways. In the most extreme cases in which the item is undiagnostic (i.e., the extent of a language model's bias is unrelated to how it will respond to this item), benchmark creators included an unhelpful element of randomness into their measure (akin to flipping a coin and changing the language model's "bias score" up or down, depending on whether the coin comes up heads). But even in less extreme cases, where the item is weakly diagnostic (i.e., the model's response to the item is a bit indicative of its bias), including such an item is detrimental: It decreases the influence that a higher quality item has on the aggregate test score. For example, for scores like "percentage of test items where the language model gave a biased response" a poor quality item is weighed equally to a high-quality one; including additional low-quality items hence reduces the

should be taken into consideration, when deciding on which items to include in a final dataset.

impact of the high-quality ones on the total score. Consequently, it is important to analyze the quality of test items (e.g., to discard low quality items).

Redundant items Unfortunately, including high-quality items does not suffice to create a high-quality test. Why that is the case becomes apparent from the following thought experiment: Imagine a highschool math aptitude test of 100 items of which 90 involve the multiplication of two-digit multiplicands (e.g., 13x24, 42x12, etc.). Despite being made of high-quality items (correct and incorrect responses clearly relate to how good students are at math), the test is only diagnostic for students for whom such multiplications are challenging but possible. The situation is different for other sections of the math aptitude spectrum: Students that cannot yet multiply will almost certainly answer incorrectly on all these items (ranging – in the ideal scenario where the 10 remaining items are easier than such multiplications – from 0 to 10 correct answers on the exam) while highly proficient students (for whom such multiplications are trivial) will likely answer them all correctly. For either of these groups of students at most the remaining 10 test items are truly diagnostic. In this situation, scores mislead: Total scores of 2 vs 4 (or 92 vs 94) appear similar, despite potentially reflecting large differences in math aptitude (as the students had 2 vs 4 correct answers on the 10 question on which they could conceivably differ).[4]

Similarly, bias benchmark items could differ: For some items, only the most biased language models would provide a biased response, while others pick up on subtler manifestations of bias (i.e., both somewhat biased and highly biased language models provide biased responses to this item). Even if it only includes high quality items (e.g., answers on them are fully indicative of how biased the model is), if a benchmark dataset – unbeknownst to us – is specialized in a particular severity of bias, it could leave misleading impressions. For example, a language model providing non-biased responses to 94% of test items – which could easily be interpreted as the model being (largely) unbiased – could simply stem from the majority of test items not sufficiently probing for subtle biases. Such misleading impressions could cause harm to ordinary citizens by legitimizing (to politicians or industry leaders) the implementation of biased language models. Relatedly, Parmar et al. (2023) recently found that benchmark datasets for Natural Language Understanding tasks lead to overly optimistic impressions about language models' abilities, because the instructions and examples given to laypeople, who generate these test items, lead them to mostly generate similar items. It will be important to ensure that we do not similarly underestimate a language model's bias, due to the items of bias benchmarks being too similar.

[4]While we write about point totals here, the same issues also arise for accuracy (a common performance metric for NLP benchmarks) or other measures of task performance (e.g., precision or recall) that are directly based on how many items a testtaker answers correctly.

Fortunately, tools exist for assessing the quality of test items (see Section 3.4) and for assessing the composition of tests (see Section 3.3.3). Specifically, in this paper, we argue for adapting evaluation techniques from the psychometric framework of *item response theory* (IRT), a framework that has been used for many years to improve tests in educational science and psychology. We hope that the application of IRT will help bias benchmarks develop towards their potential as trustworthy indicators of the extent to which a language model would cause harm to ordinary citizens, were it implemented.

3.3 Assessing the severities of model bias that a benchmark dataset can detect

Following our introduction to bias benchmark datasets, this section provides a conceptual overview of item response theory (or "modern test theory"; Crocker & Algina, 1986) and some of its tools which we believe to be valuable in the generation, evaluation and improvement of bias benchmark datasets. Broadly, item response theory (IRT; Cai et al., 2016; Debelak et al., 2022; Lord, 1980; Paek & Cole, 2019) is a theoretical framework that seeks to describe testtakers' performances on test items in terms of attributes of the testtaker (e.g., a language model's gender bias) and attributes of the test (e.g., WinoBias).

There are several different mathematical models that are based on this theoretical framework. They differ in their assumptions (e.g., about the relationship between testtakers' internal attributes and their responses), complexity, and scope (i.e., what they can and cannot account for). In this section, we will provide an overview of the general logic underlying all these models (Section 3.3.1). Subsequently, we will introduce one such model (the 1-Parameter Logistic Model; Section 3.3.2), and visualizations that this model enables (Section 3.3.3). With this methodological tool kit, we can evaluate the aforementioned (Section 3.2.4) concern about the breadth of model bias severities that bias benchmark datasets can evaluate.

3.3.1 Main concepts of item response theory

A central idea of item response theory is that a testtaker's – here: language model's – response on a test item is determined by properties of the testtaker and properties of the item. How they conceptualize and model the relationship between these item and testtaker properties is how IRT models differ (De Ayala, 2013). Broadly, differences arise from the types of test items the IRT models address (e.g., binary, or ordinal data), the shapes of their probability functions (e.g., with which mathematical formula the relationship between testtakers' attributes and answers is modeled), the (number and types of) item properties they assess (see e.g., Section 3.4), and the way they model testtakers' performances (e.g.,

whether some testtakers' performances cluster; see our discussion of hierarchical IRT models in Section 3.7).

Generally, we provide IRT models with data on how different testtakers answered on different test items (e.g., language model A gave biased responses to test items 1, 3 and 5 but unbiased ones for items 2 and 4; language model B gave biased responses to all items; etc.) and IRT models use these to estimate values for parameters that describe the testtakers and the items. At a minimum, IRT models estimate values for one such parameter per testtaker and one per item.

On the side of the testtaker, the IRT model estimates values of at least one attribute – usually called *ability* or *trait* – that causes their responses on test items. The higher a testtaker's trait level, the more likely they are to answer "correctly" on items that measure this trait. In IRT, "correctly" is shorthand for "as we would expect testtakers to respond, if they possess much of this trait". When assessing a language model's bias, "correct" responses hence counter-intuitively refers to responses that indicate that the language model is biased (as high trait levels indicate high language model bias). We will hence from now on speak of "biased" instead of "correct" responses.

On the side of the items, IRT models estimate at least the so-called *difficulty* of an item: For high difficulty items, only highly biased language models are likely to give biased responses, on lower difficulty items also less biased language models may provide biased responses. Unlike in the context of exams, where the word "difficulty" can be intuitively applied (e.g., people give fewer correct responses to difficult vs easy math items), in the context of language model bias, high or low difficulty instead refers to how blatantly a test item assesses model bias: Only highly biased language models give biased responses to test items that blatantly assess bias (i.e., on "high difficulty" items). In the following sections, we will thus speak of the "blatancy" of NLP bias test items, rather than of their "difficulty".

The more biased a language model (in IRT lingo: the higher the trait level of the testtaker) and/or the less blatant the test item, the higher the probability that the language model will give a biased response on this test item. Figure 3.1 depicts the probability of giving a biased response on a hypothetical test item: The more biased (blue) language model has a higher probability of giving a biased response than the other (orange) model. Similarly, if the test item had lower blatancy (see the dashed item curve), both language models would be more likely to give biased responses. How exactly a testtaker's value on a trait maps onto their percentage of answering an item "correctly" is defined through the IRT model's probability function. Amongst IRT models one commonly chosen probability function – which also is the basis of Figure 3.1 – is the logistic function (with normal-ogive formulations of IRT models being popular for Bayesian estimation; see e.g., Fox, 2010). In the following section, we will explain the model underlying Figure 3.1.

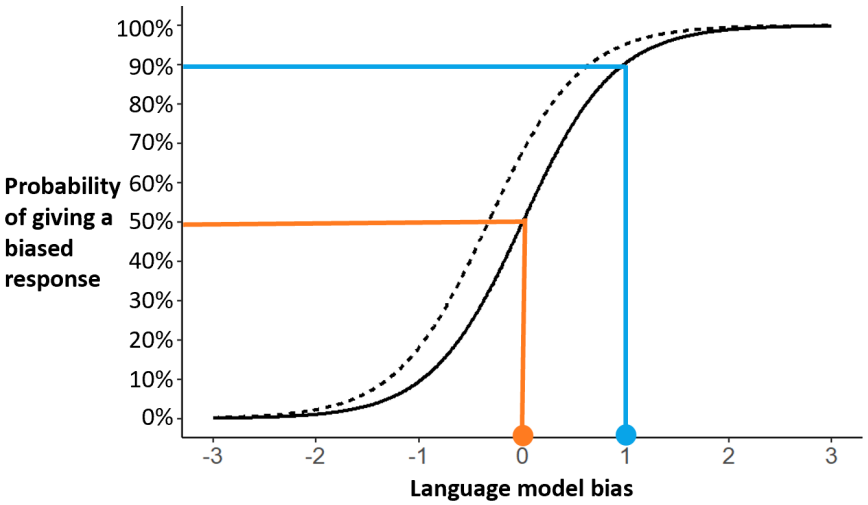


Figure 3.1: These *item characteristic curves* depict a more (solid curve) vs a less (dashed curve) blatant bias test item. A language models' level of model bias (indicated by their positions on the x-axis) influences their probability of giving a biased response on a test item (indicated by the height of the test item's curve at the language model's x-coordinate). For example, the more biased blue language model has a 90% chance of giving a biased response on the blatant item (solid curve), while the less biased orange language model only has a 50% chance.

3.3.2 The 1-Parameter Logistic Model

Figure 3.1 was created based on the simplest logistic IRT model, the so-called "1-Parameter Logistic Model" or "1PL model". This IRT model owes its name to only having one modeled item parameter, the item blatancy (or "item difficulty" in IRT lingo) b. The following formula expresses the probability function of this IRT model:

$$P(X = 1|\theta, b) = \frac{e^{\theta - b}}{1 + e^{\theta - b}} \tag{3.1}$$

This formula indicates the 1PL model's prediction for the probability that a language model gives a biased ($X = 1$) response to an item, given the language model's level of bias (i.e., level on the trait θ) and the blatancy of the item (b). If the language model's bias level equals the item's blatancy, the expression $e^{\theta - b}$ reduces to 1; the 1PL model predicts that the language model has a 50% chance of giving a biased response to this test item ($\frac{1}{1+1}$). If the language model's bias exceeds the items' blatancy, the expression $e^{\theta - b}$ grows larger (e.g., $P(X = 1|\theta = 3.5, b = -1.5) = \frac{e^5}{1+e^5}$) and the fraction approaches 1 (i.e., the language model's chance of giving a biased response approaches 100%). Conversely, the more the items' blatancy exceeds the language model's bias level, the more the fraction – and consequently the chance of the language model giving a biased response

to this item – approaches 0 (as the numerator, $e^{\theta-b}$, approaches 0 with smaller exponents). Based on the responses of all testtakers (i.e., language models) to all test items (i.e., whether they provided a biased or an unbiased response), IRT models jointly estimate each testtaker's trait level as well as each test item's blatancy level.

Numerically, trait levels are real numbers centered around 0 – with 0 representing the average in the sample of testtakers, and positive (negative) numbers expressing above (below) average levels of the trait (e.g., of language model bias). As item blatancy – in 1PL models and in section 3.4.1's 2PL models – is defined as the bias (trait level) a language model must possess to have a 50% chance of providing a biased response on an item, item blatancy and level of model bias use the same scale. This makes a language model's bias level and an item's value on the blatancy parameter directly comparable: For these IRT models, language models with bias levels above (below) an item's blatancy have an above (below) 50% chance of giving a biased response to the item.[5] The conventional unit of expressing trait levels (and hence also for item difficulty/blatancy values) are standard deviations: For example, a trait level of $\theta = -1$ indicates that the language model's estimated level of bias is one standard deviation lower than the average bias among the assessed language models.

If its model assumptions hold, even an IRT models as conceptually simple as the 1PL model can provide some information about the quality of individual items. This allows test developers to improve a test's quality – by rephrasing and reassessing items, or by removing unsalvageable ones. The quality of individual items can be evaluated visually by comparing expected item characteristic curves (i.e., the probability curve an item should have, given the IRT model's estimate for the item's blatancy) with actual observed proportions of biased responding from language models with different levels of model bias.

Figure 3.2 provides two hypothetical examples of bad items that can be detected this way. Whether language models have much or little bias barely influences their probability of giving a biased response to the item on the left. This type of item deficiency can be formally analyzed with the 2-Parameter Logistic Model (see Section 3.4.1). Other deficiencies can only be found through visual inspection of item characteristic curves. For example, the item on the right of Figure 3.2 is poorly described by a logistic function (e.g., unlike what the IRT model expects, a language model's probability of giving a biased response does not monotonically increase with increasing model bias levels). Besides assessing the quality of individual items, IRT models are also useful, because they help us explore the composition of the analyzed test. The following section will be dedicated to these kinds of explorations.

[5] Another way of conceptualizing the blatancy parameter in 1- and 2PL models is hence to think of it as a threshold: It indicates the amount of bias a language model needs to surpass for its probability of giving a biased response, $P(X = 1)$, to exceed its probability of giving an unbiased response, $P(X = 0)$.

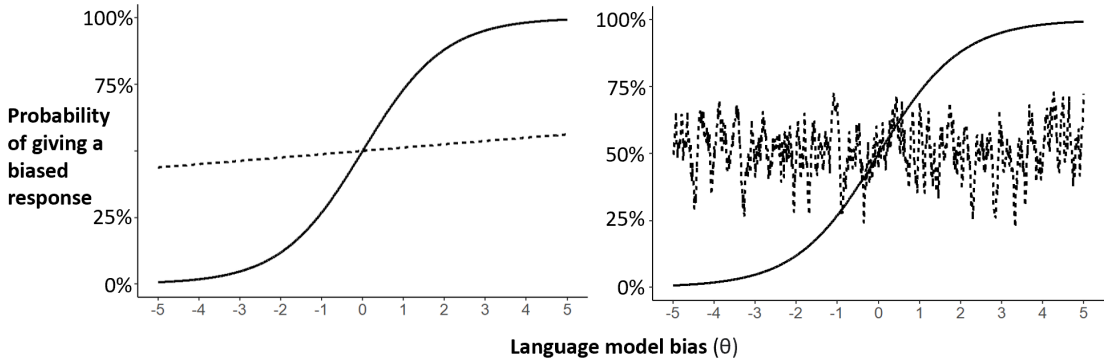


Figure 3.2: Two example item characteristic curves of low quality items. The dashed lines depict observed proportions of biased responses from language models of different levels of model bias (e.g., for the item on the left, 50% of language models with bias level $\theta = 0$ gave a biased response). The solid lines depict the item characteristic curves that the 1PL model predicts for these items, given their estimated blatancies. For the item on the left, there seems to only be a weak relationship between a language model's bias and the way it responds to this item. For the item on the right, it is uncertain whether any such relationship between model bias and model responses exists.

3.3.3 Item information and test information curves

Two useful tools that are derived from the probability functions of IRT models (like the 1PL model) are item information curves and test information curves. These two types of graphs allow us to investigate the question that we posed in Section 3.2.4: "Which bias severities can a bias benchmark dataset assess?". As we had mentioned in the thought experiment of that section, test items that are substantially too difficult (akin to asking first-graders to compute $\sqrt{24.6}$) or too easy (e.g., asking a fifth-grader to compute 1 + 2) for a testtaker provide us with little information about the testtaker's trait level: The heavily outmatched testtakers (e.g., orange arrow in Figure 3.3) will (almost) never answer these items correctly, while the insufficiently challenged testtakers (e.g., blue arrow in Figure 3.3) will (almost) always answer them correctly.

Similarly, a bias benchmark item tells us little about a language model's bias, if its blatancy is much higher or much lower than the language model's bias trait level. Instead, we learn more about a language model's bias level, the closer it is to the item's blatancy level (i.e., whether the language model provides a biased or unbiased response to this item provides us with a lot of information about its level of bias). This intuition is captured in so-called *item information curves* (IICs; see the dashed curves in Figure 3.4) which project how informative an item is at different trait levels of testtakers, given the parameter values that the IRT model estimated for the item.

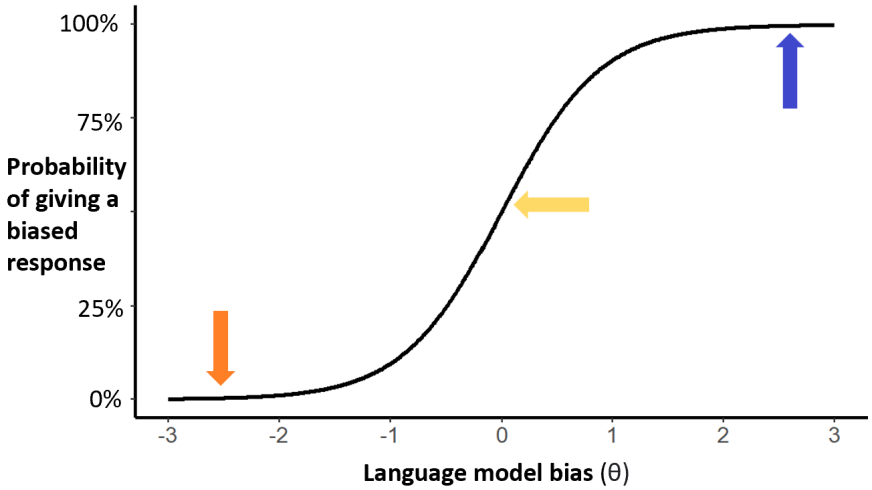


Figure 3.3: Annotated item characteristic curve for an item of blatancy $b = 0$. The item provides little information about language models whose trait levels are distant from the item blatancy (e.g., models of $\theta > 2.5$ or $\theta < 2.5$; see the blue and orange arrows). Most information is provided for language models whose model bias level equals the item's blatancy (i.e., $\theta = 0$; see the yellow arrow's tip).

For 1PL models where items only differ in blatancy, item information curves have identical (in width and height) symmetrical shapes that only differ in location (i.e., their peak is at the item's blatancy, as that is the trait level at which the item provides the most information). The information an item provides increases with proximity to its blatancy level and the distributions are symmetrical around this blatancy level (e.g., an item of blatancy $b = 0.1$ provides as much information for language models with $\theta = 0$ as it does for language models with $\theta = 0.2$).[6]

Besides providing us with information about individual items, item information curves are useful because they can be "added up" into a *test information curve* (TIC; the solid line in Figure 3.4). A test information curve provides an overview of the ranges of trait levels that a test is (most) informative about. These curves inform us about strengths (trait level intervals that are well-covered) and blindspots of our tests: For trait level ranges for which we have little test information (e.g., $\theta \geq 4$ in Figure 3.4), the IRT model's estimate for the testtaker's trait level is much less trustworthy.[7]

[6]This alignment between an item's blatancy and the θ at which it provides the most information, and this symmetry of information around an item's blatancy also holds for the slightly more complex 2PL model (see Figure 3.5), but usually not for even more complex IRT models (e.g., the 3PL model, see Figure 3.6).

[7]If you are interested in the technical details behind calculating an item's "information" at a particular trait level – the mathematical building blocks of item information curves – and of

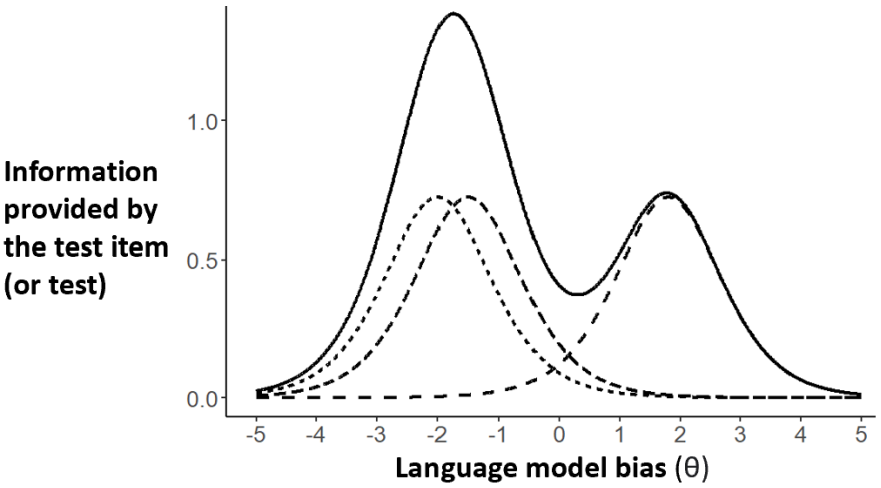


Figure 3.4: Examples of *item information curves* for 3 items (the 3 dashed lines; items of $b_1 = -2, b_2 = -1.5, b_3 = 1.8$) as well as the *test information curve* for a test consisting of these 3 items (solid line). As two items have blatancies close to -1.75, the test provides most information for language model bias of that trait level. As no items of very high or low blatancies are included, the test provides little information about language models with biases of $\theta < -4$ or $\theta > 4$.

The question of "What severities of bias can be detected with our bias measure?" can hence be answered by inspecting NLP bias benchmark datasets with test information curves. We believe that such an evaluation is very important: Given the large size of bias benchmark datasets, and given the ways in which their test items are commonly generated (e.g., eliciting items from online crowds without rewarding or otherwise ensuring the generation of items of differing difficulties/blatancies), we suspect that many items within bias benchmarks are redundant[8] (i.e., of similar blatancy, resulting in test information curves with pronounced peaks). Were that the case, our bias benchmark datasets would, unbeknownst to us, be specialized and informative only about a particular range of model bias severities.

In a more general diagnostic test, where the goal is to discriminate between language models of a range of different bias severities, we desire a less peaked test information curve – one that indicates that the test provides much information

how item information curves are combined into test information curves, we recommend the IRT primer of Warm (1978). For practical purposes, we do not consider this knowledge vital, since most freely accessible IRT software (e.g., the R package MIRT; Chalmers, 2012) graph these curves for you, and since their interpretation is intuitive.

[8]Such redundancy was already demonstrated for some non-bias NLP benchmark datasets: Polo et al. (2024) could accurately predict language models' performance on entire benchmarks based on a fraction of their items (e.g., based on only 180 of the around 29000 items from HuggingFace's Open LLM Leaderboard; Beeching et al., 2023).

(and to a similar extent) about all relevant and currently assessable[9] levels of model bias. One of the strengths of language models as testtakers is that they (unlike human testtakers) can answer thousands of questions without fatigue or motivational effects. It should thus in principle be possible to generate NLP bias measures that provide a lot of information at every (currently assessable) bias level.

Blindly generating test items without formally analyzing their difficulty level is not the way towards that goal, however. In their analysis of items from NLP benchmark datasets, Rodriguez et al. (2022) found that the IRT difficulty parameters of items clustered based on characteristics of the task (e.g., the sentiment a language model had to detect in a text segment: "positive", "neutral", or "negative"), regardless of which benchmark dataset they stemmed from. While they did not assess bias benchmarks, a similar problem could arise, here: As many popular bias benchmarks (e.g., WinoBias; Zhao et al., 2018) use singular question formats and tasks across all test items, we suspect that their items similarly cluster in regards to their blatancy levels (and hence in the levels of model bias they can detect). This is why we advocate for the evaluation of existing bias benchmark datasets with IRT models.

While these evaluations will certainly involve test information curves, they will likely require more complex IRT models than the 1PL model (e.g., because of bias benchmarks' common response formats; see Section 3.4.2). In the next section, we provide a short introduction to a few of these more complex IRT models.

3.4 More complex IRT models

Following our discussion of the 1-Parameter Logistic Model, we will in this section introduce extensions of that model which can account for additional properties of test items. The resulting, extended IRT models are named after the number of item parameters. For example, we first discuss the 2-Parameter Logistic Model (2PL; Section 3.4.1), which models two parameters: the blatancy of a test item as well as how efficiently it differentiates between language models of different bias severities. Subsequently, we discuss the 3-Parameter Logistic Model (3.4.2) which additionally accounts for performance floors (e.g., because a testtaker, when picking between answer alternatives, could randomly pick the correct response).

[9]IRT models cannot estimate the parameters of test items that either no or all testtakers answered correctly (or the ability levels of testtakers that answer all or no items correctly). In such cases the model estimates that the item has infinitely high (if there are no correct responses) or infinitely low (if there are no incorrect ones) difficulty and the discrimination value (discussed in Section 3.4) is 0, as the item cannot distinguish between testtakers. Hence, IRT models can only estimate the parameters of bias benchmark items that at least one current language model can give unbiased responses to.

These additional item parameters influence the shape of items' characteristic curves, as well as the shape of their information curves. Modeling additional item parameters should improve our judgements about the test items and the test as a whole (e.g., by influencing the shape of the test information curve). However, the additional parameters can also make the model too complicated. Parameter values are highly dependent on the IRT model that is fit. If an inappropriate model is fit (e.g., a 2PL model for test items where there is substantial probability of testtakers randomly guessing correctly; see Section 3.4.2), the estimated values can mislead. It is therefore important to first select the appropriate IRT model, before interpreting item parameter values. To do so, usually, multiple IRT models are fit to the same data. Then statistical tests determine whether the introduction of an additional parameter can be justified or whether it explains an insufficient amount of data – beyond what the model without this parameter can explain – to be included.

3.4.1 Assessing the quality of test items with the 2PL model

Compared to the 1PL model, the 2-Parameter Logistic Model (2PL model) adds the new parameter a which represents the extent to which an item can differentiate between testtakers' trait levels. The probability function of the 2PL model looks as follows:

$$P(X = 1|\theta, a, b) = \frac{e^{a(\theta-b)}}{1 + e^{a(\theta-b)}} \tag{3.2}$$

As it does for the item difficulty b, the IRT model estimates a value for this new *discrimination parameter* for each of the test items. For small values of a ($0 < a < 1$), the influence of the difference θ - b (i.e., between a language model's bias and the item's blatancy) in determining the probability of a biased response is reduced. As parameter a approaches 0, the item characteristic curve becomes flatter (see the left graph of Figure 3.5); the item becomes worse at differentiating between testtakers' trait levels (in the extreme case of $a = 0$, the fraction reduces to $\frac{1}{1+1}$ – no matter the trait level of the testtaker, they will always have a 50% chance of providing a biased response to this item).

For $a > 1$, the influence of the difference is amplified (i.e., a language model's bias relative to the item's blatancy determines the probability of a biased response to the item, more). Items with larger values for a have steeper item characteristic curves (and hence more peaked item information curves; see Figure 3.5): The item is better at distinguishing between testtakers with trait levels close to its difficulty (or here: blatancy), but at a narrower interval around that difficulty (e.g., compare the purple with the blue IIC in Figure 3.5).

The estimation of 2PL parameters for test items thus allows us to evaluate their quality in more detail: Items with larger as are better at discriminating

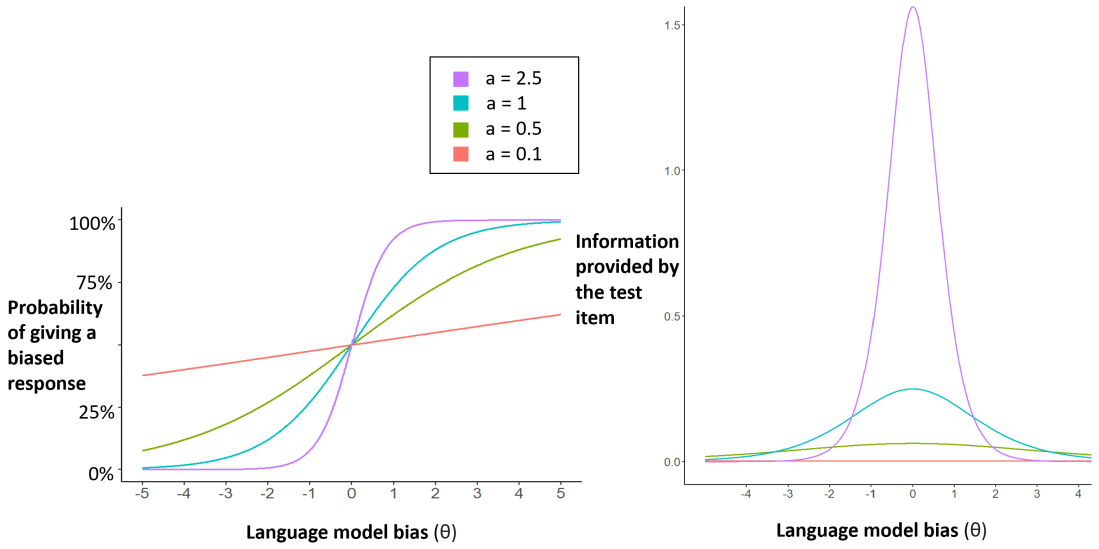


Figure 3.5: Item characteristic curves (left) and item information curves (right) of items with blatancy $b = 0$ that differ in discrimination parameter values. Items with large values for a discriminate better between testtakers whose θs are close to the item's blatancy (left figure). Consequently, the item provides more information for language models of such trait levels (right figure).

testtaker's ability levels (e.g., language model's bias), items with low as are often better revised (and reassessed) or discarded. An ideal bias measure would consist of items with high discrimination values whose difficulty levels are spread across all relevant (and currently assessable) levels of model bias, so that the bias measure provides information about all these levels. This ideal situation is hard to reach in practice – item generation is difficult (especially at extreme trait levels, like very small model bias) and in NLP, small changes to an item's phrasing can significantly alter how language models answer them – but analyzing the quality of test items is critical. Assessing their items allows us to iteratively improve bias measures (e.g., by deleting or rephrasing the worst performing items and by subsequently reevaluating the bias measure) and to communicate their limitations which, in absence of IRT analyses, could go unnoticed (e.g., that our bias measure only provides reliable information about a particular range of model bias severities).

In sections 3.2.4 and 3.3.3, we discussed that items' blatancies influence the extent to which count-based evaluation metrics (like point totals or accuracy) can be interpreted as indexes of how biased a language model is (e.g., if most test items are of high blatancy, a low percentage of biased responses is expected). Here, we introduce another risk of not assessing items' IRT parameters: Adding items of low a also decreases the extent to which point totals are a valid index, because point totals weigh low and high a items equally. By taking into account

more information about an item than merely whether a testtaker answered it correctly, an IRT model's estimate of a testtaker's trait level addresses some of the weaknesses of count-based evaluation metrics; a discussion we return to in Section 3.5.1.

3.4.2 Accounting for response formats with the 3PL model

Compared to the 2PL, the 3-Parameter Logistic Model (3PL model) introduces a new parameter c which, for each test item, estimates a performance floor. The probability function of the 3PL looks as follows:

$$P(X = 1|\theta, a, b, c) = c + (1 - c)\frac{e^{a(\theta-b)}}{1 + e^{a(\theta-b)}} \tag{3.3}$$

For example, for items that have an estimated parameter value of $c = 0.50$, the fraction $\frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}}$ now only explains half (i.e., $1 - c$) of the probability of giving a biased response and language model's probabilities can range from the performance floor of 50% up to 100% (instead of 2PL's 0-100%). With the change in possible probability values, the interpretation of item blatancies changes: They now express the trait level (θ) testtakers need to split the remaining amount of total explainable probability in half. Instead of indicating a 50% chance of biased responses, for $c = 0.50$ blatancy indicates the level of bias (θ) at which language models have a $50 + 50/2 = 75\%$ chance of giving a biased response.

For each test item, a value of c is estimated. A non-zero value of this parameter influences the shape of the item's item characteristic curve and item information curve (see Figure 3.6) and improves the extent to which the item's other parameter values (e.g., for item blatancy) are meaningful. To understand why that is the case, consider a difficult multiple choice math item with two possible answer options. If testtakers were to guess randomly, they would – independently of their trait level – have a 1/2 chance of answering correctly. A 2PL model would not account for this chance of randomly choosing the correct option, and would (misleadingly) assign a very low item difficulty to this item (as item difficulty in the 2PL is defined as the trait level at which the testtaker has a 50% chance of answering correctly).

As the parameter c accounts for any probability of arriving at a correct (or here: biased) response that is not due to the trait level θ, it is referred to as the *guessing parameter* or the *pseudo-guessing parameter*. The "guessing" part of the name comes from random guessing, which is one common avenue through which human testtakers can give correct responses to items, independently of their trait levels. The "pseudo-" indicates that the parameter c may accommodate not only guessing but also other potential causes for a performance floor. While the insentient language models cannot be said to "guess", we believe that this parameter is highly relevant to the evaluation of bias benchmark response data: Many prominent bias benchmark datasets diagnose bias through relative

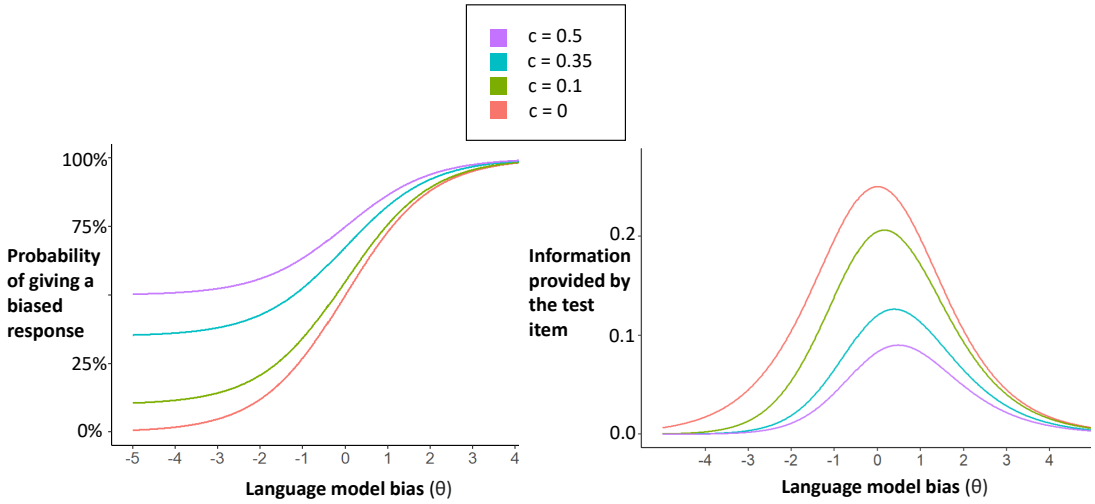


Figure 3.6: Item characteristic curves and item information curves of test items that differ in the pseudo-guessing parameter c, but have identical values for blatancy ($b = 0$) and discrimination parameter ($a = 1$). With larger values of the pseudo-guessing parameter, there are fewer response differences between testtakers that an item can model (e.g., across trait levels, the purple item approaches the 100% ceiling sooner than other items; left figure). Consequently, items with larger values of c are less informative (right figure). Additionally, the parameter leads to shifts in the information distribution: A $c \neq 0$ shifts the mode of the information distribution (i.e., the θ at which the item is maximally informative) to a value larger than the item's blatancy and skews the information distribution negatively (i.e., shifts it so that more information is provided for bias levels above the mode than for bias levels below it). The larger the value of c, the more pronounced these shifts.

performance or relative preference and thereby force a language model to choose from a restricted number of potential options. When, for example, comparing semantic similarities (e.g., the similarities between "the doctor walked home" on the one hand, and "the man walked home" vs "the woman walked home" on the other hand), the language model necessarily (e.g., already because of measurement error) will have a preference for either the stereotype-consistent or the stereotype-inconsistent sentence pairing. For items with such answer formats (i.e., forcing or prompting the language model to choose one of two options), we would expect that the 3PL model estimates values of $c \approx 0.5$, because a priori – even if it had virtually no language capacity – the language model would have a 50% chance of randomly making the stereotype-consistent ("biased") choice.

For such types of items – and therefore for many bias benchmark items – 3PL models likely give us better estimates of testtakers' trait levels (e.g., of language model bias) and of items' parameters (e.g., item discriminability), because they

consider the a priori odds of language models providing different responses. On the other hand, 3PL models require more data (e.g., more language models' responses to bias items) than IRT models that estimate fewer parameters per item. As testing a large number of language models is hard (and one commonly requires hundreds of testtakers to adequately fit 3PL models; Akour & AL-Omari, 2013), an alternative approach to making performance floor adjustments might be necessary in the NLP bias case: Instead of fully fitting a 3PL model and hence estimating a pseudo-guessing parameter value per test item, one can also opt to fit (the equivalent of) a 2PL model with a (theoretically-motivated) lower asymptote for each relevant item (e.g., a 33% performance floor for a 3-option multiple choice item). This way, adjustments for performance floors – and hence adjustments for the common response formats of bias benchmark items – can be made without requiring as much data as do full 3PL models.

3.5 Consequences of IRT for NLP model bias

Following our more conceptual description of popular IRT models and their parameters, we dedicate this section to the application of these concepts. On the theoretical side, we discuss the implications of IRT for the interpretation of common NLP bias evaluation metrics like differences in accuracy (see Section 3.5.1).

On the practical side, we argue for a large-scale evaluation of existing NLP bias benchmark datasets and discuss how one of IRT's large downsides for the application to the NLP bias case (i.e., IRT models' demands for data from many testtakers) could be addressed (see Section 3.5.2).

3.5.1 Accuracy-like evaluation metrics are precarious

Most (if not all) NLP bias benchmark datasets are evaluated with metrics derived from proportions of language model responses that correspond to target labels (e.g., accuracy, precision, recall, F_β scores). For instance, WinoGender (Rudinger et al., 2018) uses differences in accuracy (the percentage of items that a language model answers correctly) to assess differential performance on stereotype-consistent versus stereotype-inconsistent sentences. That is: It tests whether the language model is more accurate on stereotype-consistent sentences. Similarly, WinoBias (Zhao et al., 2018) assesses performance differences (through differences in F1 scores) between "pro-stereotypical" and "anti-stereotyped" sentences. Also performance metrics that were specifically designed for bias benchmarks (e.g., s_{AMB} from the BBQ benchmark; Parrish et al., 2021) are usually based on such proportions of correctly labeled model responses.

However – if we have data from a sufficiently large number of language models to fit an appropriate IRT model (see Section 3.5.2) – an IRT model's estimate of the language model's trait level is usually a better evaluation metric. Like

these count-based metrics, trait estimates take into account whether items are answered correctly or incorrectly, but they also consider other information about the items and hence take into account that "not all items are created equal".

Item blatancy / item difficulty Any IRT model takes into account the *difficulty* (or here: "blatancy") of different test items. Including redundant items (e.g., many items of similar blatancy) influences accuracies[10] immensely: If a large proportion of test items are of similar blatancy, small differences in language model bias (e.g., $\theta_a = -0.5$ vs $\theta_b = 0$, if most items have a blatancy of $b = -0.25$) can lead to large differences in observed accuracies. IRT models' θ estimates are more informative. For example, a language model (e.g., with $\theta = 2$) giving biased responses to a large number of redundant low blatancy items (e.g., items with $b = -0.25$) will not sway the IRT model's estimate of the language model's trait level much, as language models of high bias θ are expected to provide biased responses on these items.

Item discriminability The values of the *discrimination parameter* (for models like the 2PL or 3PL) of items also influences how we should interpret a language model's responses to items. On items where there are few differences between how language models of low vs high bias levels respond (i.e., on items with low values for a; see Section 3.4.1), we should put little weight on whether or not a language model provides a biased response. Conversely, items with high discrimination values are highly indicative of the language model's trait levels (i.e., levels of bias) and should hence be taken into account more. IRT models can consider these qualities of test items; count-based measures like accuracy treat all items as though they were the same. By treating performance as probabilistic[11] and dependent on discriminability, trait level and blatancy, IRT models have the additional benefit of accounting for (some) misleading accuracy scores – cases in which a more biased language model provides fewer biased responses to test items than does a less biased language model[12].

[10] In this section, we use the conceptually simpler accuracy as a stand-in for other count-based evaluation metrics like (differences in) F_β scores, or s_{AMB}. While some of these are arguably conceptually superior to accuracy (e.g., F1 scores are more informative for unbalanced labels), none of them account for the attributes of test items. In other words: They may have advantages over accuracy, but – compared to IRT models' trait estimates – share the same weaknesses.

[11] By that we mean "we do not a priori know whether a language model will give a biased response to an item that it had never been previously subjected to" (though, based on its estimated trait level, we know how likely it is to give biased responses on items of this blatancy). Of course, for several (types of) language models, responses to an item are deterministic in the sense that they will always provide the same answer to the same item.

[12] While counterintuitive, the fact that IRT models can estimate higher trait levels for language models that provided fewer biased responses to test items is desirable: Firstly, there are some test items for which a biased responses is so worrying, that an unbiased response on another item is not equally reassuring (e.g., a language model gave a biased response on an

Response formats and (pseudo-)guessing The fact that language models (in several response formats) have high a priori probabilities of giving certain responses (e.g., that there is a 50% chance of picking the stereotype-consistent pronoun, when choosing between "she" or "he") is another feature that dampens the extent to which accuracies (or accuracy differences) are meaningful: A 70% accuracy might seem like a decent result but is much less impressive if an accuracy of 50% is already expected just by chance. IRT models (which include a pseudo-guessing parameter) can estimate and provide metrics that are comparable across answer formats (e.g., enabling the comparison of different bias benchmark datasets): Whether or not there are reasons to suspect performance floors on an item (e.g., if an IRT model simultaneously assesses items with and without such a floor), the interpretation of item blatancy remains the same: Positive (negative) b values indicate items of above (below) average blatancy, compared to all other items that were analyzed by this IRT model. Similarly, estimated trait levels can be interpreted, independently of a test's response format: Positive (negative) trait levels indicate that the language model had a higher (lower) than average level of bias, compared to all other language models whose answers were analyzed by this IRT model.

Additional context On top of a performance metric (i.e., estimated trait level), IRT models provide additional important context for interpreting said metric: Firstly, IRT models do not only provide an estimate for a language model's level of bias but also standard errors for the trait level estimates. These standard errors (and standard error-derived *confidence intervals*) are ways of representing uncertainty about a language model's level of bias (see e.g., Ethayarajh, 2020, for a discussion of why drawing conclusions about NLP models' biases without such measures of uncertainty can be problematic).

Secondly, we can explore through item characteristic curves whether test items function similarly for different subgroups of language models (e.g., subgroups based on different model architectures). When the assumption of similar item behavior across testtaker subgroups does not hold, we speak of *differential item functioning* (DIF; see Figure 3.7). Testing bias benchmark items for DIF will be important to assess the extent to which bias benchmark items can be universally applied (and the extent to which scores can be directly compared) across different types of language models (or, e.g., whether a biased response on a particular benchmark item is more indicative of bias for one type of language model than for another). If DIF proves common for bias benchmark items, it might be necessary to generate and validate separate bias measurement tools for different subpopula-

item of blatancy 3, and an unbiased one to an item of blatancy 1). Secondly, some items are of higher quality (e.g., discrimination value) than others. If a test, for example, had 3 high-quality items and 27 coin tosses (i.e., 27 test items with discrimination values ≈ 0), it would not be surprising that some testtakers with low trait levels have a higher total score than do some testtakers with high trait levels.

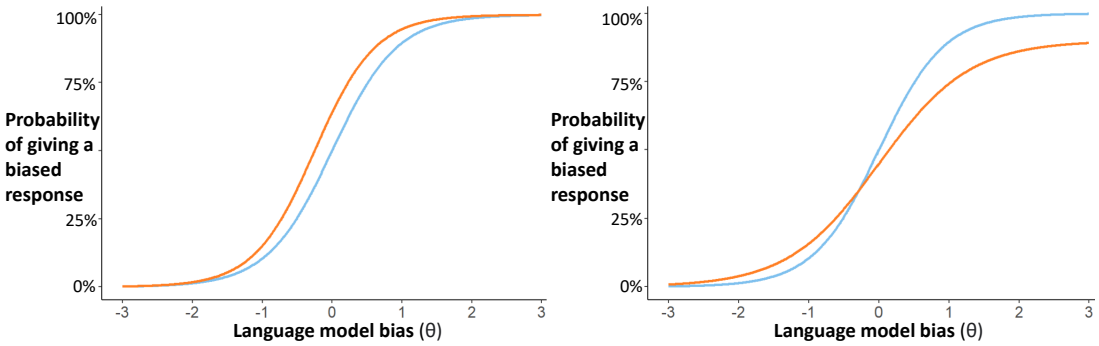


Figure 3.7: Each figure depicts hypothetical item characteristic curves (ICCs) that two types of language models (orange vs blue) obtain for the same bias benchmark item. In the left figure, we observe *uniform differential item functioning*: While the ICCs have (roughly) the same shape across all levels of model bias, the item is consistently more blatant for the blue than for the orange group of language models. In the right figure, we observe *nonuniform differential item functioning*: There are differences in the item's behavior, but the extent by which a group of language models is more likely to give a biased response (compared to the other group) is not consistent across bias levels (i.e., levels of θ). In this example it also differs across bias levels which of the two groups of testtakers is more likely to give a biased response (i.e., at first it is the orange group of language models, then it is the blue group).

tions of language model testtakers, since these types of language models evidently act in distinct manners.[13]

While common evaluation metrics like accuracy have the disadvantage over IRT model's θ estimates of disregarding test items' characteristics, they also have one substantial advantage: They require a lot less data and – crucially – do not require data from several different language models (see next section). Due to their data demands, it will not always be feasible to fit an IRT model and to thereby obtain (meaningful) trait level estimates for a language model. Nevertheless, we hope that our discussion of their pitfalls suggests that metrics like accuracy – in absence of additional information about the attributes of a test set – should be interpreted with caution. A language model's performance on such metrics, if interpreted relative to other models' performance, is suggestive of differences between models' trait levels, but can sometimes mislead (i.e., sometimes the model that gave fewer biased responses on the test items is more biased) and can never provide information about the magnitude of differences between trait levels (e.g.,

[13]For a more elaborate discussion of DIF and ways of detecting and mitigating it, we refer interested readers to Wu et al. (2016). On the practical side, DIF detection for all here-discussed IRT models is possible for example with Hladká and Martinková (2020)'s R package difNLR.

accuracies of 90% vs 81% are not indicative of the first model being 10% better than the second).

3.5.2 Steps to take towards better model bias benchmarks

In the previous sections we outlined how IRT models can inform us about the quality of existing bias benchmark datasets: They can help us assess the benchmark as a whole to test whether there are levels of bias blatancy that are not well-assessed by our benchmarks (see Section 3.3). Additionally, they can help us weed out bad items (see Section 3.4.1) and can provide an evaluation metric that is conceptually superior to traditional ones (see last section). Consequently, we advocate for a large-scale assessment of model bias benchmark datasets to assess their attributes (e.g., the blatancy ranges they assess) and quality, to weed out bad items and to find a subset of high quality items (across benchmarks) which can collectively serve as a high quality benchmark dataset.

When assessing bias benchmarks, one potential challenge is the dimensionality of model bias: Some authors' benchmarks (e.g., Nadeem et al., 2021) compute aggregate scores across types of stereotypes (e.g., across gender, racial, and occupational stereotypes) and hence de facto treat model bias as unidimensional (e.g., that there is one overarching "model biasedness", for which gender biased and racially biased responses are two of its manifestations). However, the jury is still out on whether types of model biases are dependent or independent (e.g., whether a language model could, in practice, have a strong gender bias but no racial bias). During the initial IRT assessment of bias benchmarks we hence advocate for analyzing items from only one type of stereotypes. This circumvents the bias dimensionality issue: whether or not this subtype of model bias exists independently of other bias types or conversely represents a more over-arching general model bias, there is only one trait that causes the responses of language models. Of the (potential) subtypes of model bias, a starting point could be (binary) gender bias measures, as that type of bias has received a lot of research attention and is assessed in many popular model bias benchmark datasets (CrowS-Pairs, Nangia et al., 2020; StereoSet, Nadeem et al., 2021; WinoBias, Zhao et al., 2018; WinoGender, Rudinger et al., 2018, i.a.).

Demand for many different testtakers One potential challenge facing the large-scale IRT assessments of benchmark datasets is IRT model's sample size demands: While recent advances in Bayesian hierarchical estimation could decrease these demands (see e.g., Gilholm et al., 2021; König et al., 2020), a reliable estimation of IRT model parameters can require data from hundreds of testtakers – numbers that are possible to reach when assessing human participants but

potentially prohibitive when assessing language models.[14]

Additionally, IRT models require testtakers of different trait levels; that is: Language models with a range of different bias levels to make IRT model estimation possible. In previous NLP studies that applied IRT to the assessment of language models' abilities, multiple approaches were used to reach this range of trait levels and to meet these sample size requirements: Lalor et al. (2016) and Fang et al. (2024) modeled the performance of human participants alongside the performance of language models on the same test items. Byrd and Srivastava (2022) trained many different instances of the same language model architecture on different subsets of training data for different training durations. Vania et al. (2021) increased their number of testtakers by testing the same models at multiple training epochs. For normal language modeling tasks, (as models tend to improve with more training) different amounts of training ensure a larger breadth of trait levels. Time will tell whether and to what extent bias similarly emerges for language models over the course of training (as bias is not deliberately taught to a model).

If we can and do apply such ways of bolstering the number of testtakers to the NLP bias case, we should be cautious about the interpretation of our results: Including many similar language models could decrease the extent to which conclusions that we draw about benchmark datasets generalize, as the analyzed models resemble each other more than they resemble other (not analyzed) language models (see e.g., also our discussion of hierarchical IRT models in Section 3.7). This is problematic, because bias benchmark datasets are supposed to be widely applicable (across model types) diagnostic tools. To ensure greater generalizability, the assessment of bias benchmarks would hence ideally involve analyzing a large variety of realistic language models (i.e., models that were trained for real-life application, not specifically for this study).

In addition to such a large scale investigation of existing and popular benchmark datasets, we also encourage developers of future datasets to fit at least simple IRT models (e.g., the 1PL model), during test development. As we outlined in Section 3.3.3, even simpler models can help with the identification of poor items (through the manual inspection of item characteristic curves) and of glaring weaknesses – like the benchmark dataset providing information about only a small range of bias blatancies. Such analyses also enable the dataset developer to be mindful and transparent about the dataset's strengths (e.g., which bias blatancies they address well) and weaknesses.

[14]Several authors (e.g., Akour & AL-Omari, 2013; Sahin & Anil, 2017) found that the number of required testtakers decreases with an increasing number of test items. In such studies, the largest tests had < 100 items. Benchmark datasets frequently have hundreds to thousands of items. Consequently, we are optimistic that the required number of different language models for fitting IRT models will be substantially lower than the required number of human testtakers.

3.6 Related work

While this article is, to the best of our knowledge, the first one to advocate for the application of item response theory to NLP bias benchmark datasets, it is neither the first paper to apply psychometric concepts to NLP bias measures nor the first to advocate for the application of IRT in NLP contexts. In the following section, we hence embed our paper in the existing literature of psychometric work on NLP bias measures, as well as in relevant NLP work utilizing item response theory.

Work on improving bias measures Following Jacobs and Wallach's (2021) seminal paper on the measurement of "algorithmic fairness" in computational systems, several authors have begun evaluating existing bias measures through the lenses of the psychometric concepts "construct validity" (roughly: the extent to which our measure assesses the concept we intend it to) and "reliability" (i.e., the extent to which measurements from a measure are consistent). For example, Du et al. (2021) and Zhang et al. (2020) investigated the extent to which gender model bias measures are reliable, while Blodgett et al. (2020) noted that there are substantial deficits and disagreements in researchers' definitions of the "model bias" concept.[15] Others (e.g., Bommasani & Liang, 2022) used these notions of validity and reliability in the creation of new bias measures.

From papers of this broader research tradition, most closely related to our paper is the work by Blodgett et al. (2021), which highlights potential issues with the operationalizations of bias in bias benchmark datasets. While they provide a qualitatively derived index of potential pitfalls that should be avoided, we provided a list of (IRT derived) quantitative techniques that can be used as diagnostic evaluation tools of benchmark datasets. We believe that these two approaches best work complimentarily: For example, to evaluate items from a benchmark dataset, instead of randomly picking a subset of items (as did Blodgett et al., 2021), an IRT analysis could identify items of interest (e.g., those with particularly high or low discrimination values or blatancies). These items could then be qualitatively evaluated (e.g., through a list of desiderata or Blodgett et al.'s inventory of common pitfalls) to identify potential reasons for why the test items performed as they did. Knowledge of which attributes best predicted high-quality items could in turn improve generation practices for future bias benchmark items.

[15] In this paper, we deliberately remained agnostic in regards to which definitions of "model bias" are appropriate or "correct" (if a singular correct definition exists). From IRT's perspective, the trait is whatever statistics identify to be the best (latent) common predictor of performance on analyzed test items. Independently of how the designers of benchmark datasets conceptualize model bias, as long as their test items are largely consistent with that conceptualization, the IRT model will likely identify their bias trait.

NLP work on IRT Item response theory is a theoretical framework that has enjoyed decades of conceptual work and research. Consequently, our discussion of IRT concepts and models in this paper could only scratch the surface of this literature. While applications of IRT concepts to NLP are rare, there are three such works that we wish to highlight, as we consider them potentially relevant to the language model bias case.

Firstly, Amidei et al. (2020) recently explored how *Graded Response Models* (a class of IRT models for ordinal data like "agree", "neutral", "disagree"; Samejima, 1969) can be used to analyze the idiosyncrasies of different human annotators. While their target application was the human evaluation of computer generated texts, we believe that a similar application of graded response models might indirectly improve the quality of future bias benchmarks: Whenever the inclusion of (e.g., laypeople-generated) test items into a bias benchmark depends on aggregate lay judgments from human annotators (e.g., if items are included into a test, whenever four of five annotators agree on the inclusion), these IRT models could, for example, point towards annotators whose approval behavior significantly differs from that of other annotators. After such identifications, benchmark dataset designers may, for example, elect to reduce the influence of overly liberal annotators on inclusion decisions. This could prevent some lower quality items from being included into bias benchmarks.

Secondly, Vania et al. (2021) recently introduced (item characteristic curve-derived) *Locally Estimated Headroom scores*, a measure of how likely benchmark datasets are to (not) be trivially easy for the next generation of state-of-the-art language models. With these scores, Vania et al. tried to address the issue of language models rapidly growing in capacity, which can make older benchmark datasets obsolete. While we are unsure about the extent to which language models "outgrow" bias benchmarks (as language models, unlike is the case for language modeling tasks like coreference resolution, are not trained with the express purpose of performing well on bias benchmarks), it is similarly important to assess the extent to which bias benchmarks can assess subtler levels of model bias than the ones current language models display.

Finally, Rodriguez et al. (2021) recently argued that traditional NLP *leaderboards* – websites where the best-performing language models' testscores on benchmark datasets are depicted – should be "re-imagined" as Bayesian IRT models. While the authors do not explicitly discuss performance on NLP bias benchmarks, some of their critiques of traditional leaderboards are relevant to performance metrics of bias benchmarks, and their discussion ties in with our critique of such metrics in Section 3.5.1.

3.7 Discussion

Given their wide application in today's connected society, assessing the extent to which language models are biased is crucial: As many ordinary citizens interact with applications (like translators) that rely on language models, even (seemingly) small undesired trends in language models' behaviors can, on aggregate, cause substantial harm. High-quality measurement tools would allow us to diagnose biased language models, before they are implemented and can cause harm. Legislators could make use of such high-quality measurement tools to implement guidelines for acceptable language model behavior. Finally, high-quality bias measurement tools are crucial because they enable systematic bias mitigation efforts (e.g., trying whether an intervention leads to a decrease in measured language model bias).

Contemporary bias measurement tools do not yet live up to the bar of "high quality" (with flaws highlighted in e.g., Blodgett et al., 2020, 2021; Goldfarb-Tarrant et al., 2021). Consequently, many attempts have been made to improve the quality of existing and newly-introduced bias measurement tools. Due to NLP's unfamiliarity with issues – like "how does one measure concepts that cannot be directly observed?" – that have long research traditions in other research fields, we believe that interdisciplinary work is best equipped to address the NLP bias case (van der Wal et al., 2024). Here, we provide such interdisciplinary work, combining NLP expertise with insights from the psychometric framework of item response theory.

In our view, a systematic application of IRT concepts to bias benchmark datasets could prove crucial. It would allow us to assess the quality of individual test items as well as the properties of the benchmark dataset as a whole. If a bias benchmark, unbeknownst to us, only assesses a small range of model bias severities, conclusions that we draw based on scores on that measure could mislead. Language models that were declared safe (e.g., because there are small differences in accuracies across stereotypical and counter-stereotypical sentences) could, in fact, be biased (e.g., because most items of the benchmark are of high blatancy) and cause harm to ordinary citizens, when implemented. A test information curve, which shows that the benchmark only provides us information about a small range of model bias severities, could make us aware of such deficiencies. Similarly, the simultaneous analysis of items across bias benchmark datasets would allow us to assemble a combined benchmark dataset that consists of high-discrimination test items (from several different bias benchmarks) which are spread across a range of different item blatancies. Finally, we argued that IRT model's trait level estimates are conceptually superior to traditional evaluation metrics (such as accuracy, F1 score, recall, etc.), because the IRT models work with more information about test items than standard measures do and because they provide us with statistics to conceptualize our (un-)certainty about the trait level estimates (i.e., standard errors and confidence intervals).

While we are enthusiastic about the prospect of applying IRT to the NLP bias, we anticipate a few potential challenges to the implementation. The first notable limitation is a conceptual one: IRT models benefit from being fit on data from testtakers with a range of different trait levels (e.g., ideally data from language models from all sections of the model biasedness distribution). In previous applications of IRT to language models, variety in trait levels could be ensured artificially: As language model capacity on tasks such as coreference resolution depends on the amount of training the models received, systematic manipulations between language models of the amount of training data they received ensured a healthy range of trait levels. The situation is less simple for the model bias case. In the absence of trustworthy model bias measures, little is known about the distribution of model bias severities (e.g., whether large and low model bias severities are equally common) or about the shifts in model bias across training time (e.g., unlike what is the case for coreference resolution, we cannot simply vary between language models the training durations and assume that this will lead to a spread in trait levels). In our view, all we can do is simultaneously assess a large variety of different language models and hope that there are no striking clusters of bias trait levels (e.g., a large number of very high and/or very low bias levels instead of the trait levels being somewhat evenly distributed). Were we to find such clusters (which itself would be a theoretically insightful finding), we would need to be careful not to overgeneralize our findings (e.g., not to make strong statements about unobserved bias severities).

A second conceptual challenge regarding the application of IRT to NLP bias is the high number of testtakers from which we require response data. How many (testtaker) language models we require data from remains to be seen. For instance, 3PL models, that allow for estimating the (pseudo-)guessing parameter, which we found of particular interest, generally require a lot of data; for example, for 60 test items, Akour and AL-Omari (2013) estimate that 500 testtakers are required to receive sufficiently precise item parameter values.[16] Additionally, common NLP ways of bolstering the number of (assessed) language models (e.g., testing the same models repeatedly at different times during training) could prove problematic: They might prove inapplicable to the bias case (depending on how bias develops across model training – a topic of current research) and they might lead to non-generalizable results (e.g., because the language models that are assessed in the IRT model are nonrepresentative of language models more generally). That being said, it is currently unclear what number of different language models will be required to achieve adequate IRT model fit – a deficit we hope to see addressed in (e.g., simulation) studies in the future. Provided that testtakers' trait levels are decently spread, previous studies have found that the required number of testtakers reduces with increasing numbers of test items that

[16]Here the aforementioned fitting of simpler IRT models with lower asymptotes could prove an important (i.e., lower-sample size) alternative to fully fitting 3PL models.

are modeled (Akour & AL-Omari, 2013; Sahin & Anil, 2017). However, these studies did not assess test sizes that resemble those of bias benchmarks (e.g., Akour and AL-Omari's largest test size was 60, while StereoSet contains more than 4000 items). Consequently, we are optimistic that the number of different language models required to fit IRT models is substantially lower than what analyses of human tests suggest.

Looking ahead Our goal was to provide an initial introduction to and conceptual discussion of basic IRT concepts that are relevant to the NLP bias case. Hence, the scope of our discussion was limited. There are several IRT concepts that we consider relevant to the NLP bias measurement case, but that would have required more statistical and conceptual discussion than we could justify here. Three such concepts that we wish to mention here are multidimensional IRT models, hierarchical IRT models, and computerized adaptive testing.

Some authors (e.g., Brown et al., 2020; D'Amour et al., 2022; Webster et al., 2021) have put forth the argument that language models become less biased with increasing language capacity (e.g., because early in training they rely on stereotypes to make predictions, while these stereotypes become obsolete once the language model gains in language capacity). More generally, a potential conceptual weakness of NLP bias benchmarks is that they are not pure measures of NLP bias, but instead assess both bias and the primary language modeling task that indirectly reveals this bias (e.g., for WinoGender: coreference resolution). This conceptual weakness could potentially be addressed and the aforementioned argument could potentially be investigated by use of multi-trait (or "multidimensional") IRT models (Bonifay, 2019; DeMars, 2013). These models allow for the simultaneous modeling of different performance-causing traits, allowing us to quantify the influence of either of these traits (e.g., model bias and coreference resolution capacity) on responses to test items.

Moreover, hierarchical IRT models (e.g., Cai et al., 2016; Rijmen, 2011) are potentially relevant, because they allow us to model dependencies between testtakers: cases in which one testtaker's answers to test items tells us something about responses of at least one other testtaker (e.g., on a national exam, a student is likely to perform somewhat similarly to their classmate, as the two share teacher and school resources). We believe that such modeling techniques are relevant to the NLP bias case, because common ways of increasing the numbers of testtaker language models (e.g., testing the same model at different times during training; Vania et al., 2021; or training many instances of the same model architecture; Byrd and Srivastava, 2022) could introduce such dependencies.

Additionally, we think that such dependencies might become even more prevalent, given current trends in NLP: Some modern language models are so powerful, that they can take on different "roles" (e.g., a chatbot can be made to answer questions as a therapist or teacher would), when provided with different sequences

of instructions and rules (called prompts; see e.g., White et al., 2023). While these (prompted) different "personas" of the language models can act substantially differently, they still stem from the same language model. If an IRT analysis of their behaviours becomes relevant (e.g., if companies implement differently prompted versions of their language model for different purposes), we would expect that differently prompted models that stem from the same language model have (somewhat) dependent observations: Independently of the prompt (e.g., be it that the model was told to act like a teacher or like a therapist), its fundamental language capacity (i.e., which language regularities the model picked up on; see Section 3.2.1) is identical.

The final IRT concept we wish to highlight here is *computerized adaptive testing* (CAT; see e.g., Klinkenberg et al., 2011; Magis et al., 2017). CAT is a method of testing that involves a sequential back-and-forth between a) estimating a testtaker's trait level (based on their performance on previous test items) and b) selecting and administering a new test item, based on the trait level estimate. For example, the testing program could select as its first test item an item that is maximally informative at $\theta = 0$ and, depending on the testtaker's answer, (re-)estimate the testtaker's trait level as slightly higher or lower than $\theta = 0$. Subsequently, the test program would choose and administer an item that is maximally informative at its new estimate of the testtaker's ability level. This process is repeated until the CAT algorithm arrives at a precise estimate of the testtaker's trait level (e.g., until the standard error of a testtaker's trait level estimate shrinks below a predetermined value).

One of the main advantages of CAT is that fewer test items are required to receive a good estimate of a testtaker's ability. For example, in a preset collection of items, it is likely that several items have substantially too high or too low difficulty (or here: blatancy) to be informative about a testtaker. Fewer such uninformative items are administered, if items are chosen specifically for a testtaker's (likely) trait level. Such efficiency gains, while less important for the assessment of individual language models (since language models can answer hundreds of items quickly, we can afford a few uninformative items), could prove relevant if we want to test a large number of models and/or if we want to repeatedly test language models, for example at different training steps (as did e.g., Lalor & Yu, 2020). We have previously argued (van der Wal et al., 2024) that such assessments across training steps are conceptually important, as they provide information about how model bias emerges and changes over the course of model training. This could provide us with hints about the causes of bias: For example, if gender bias emerges relatively early in training and then decreases in subsequent training steps, it might be that stereotypical gender associations (e.g., "nurse = female") initially represent good heuristics for task performance that later – as the model improves in the task – outlive their usefulness. We believe that CAT could play a role in making such repeated assessments of multiple language models.

More generally, we hope that our introduction to IRT concepts encourages readers to explore the large and varied literature of item response theory to find inspiration for bias measure improvement efforts – be they inspired by concepts that we introduced here, or by others. This article is meant to be a mere starting point for applying IRT concepts to the NLP bias case. While we expressed our view on how such concepts could be used towards our ultimate goal – reducing the harms that language models cause – we look forward to hearing perspectives that alternate from ours or expand beyond it.

Acknowledgments

The authors wish to thank Petr Palíšek, Alina Leidinger, and the two anonymous peer reviewers for their thoughtful and insightful feedback!

Chapter 4

Introducing OccupatiEN — a Dataset of Occupational Gender Stereotypes to validate NLP gender bias measures

Dominik Bachmann[1,2], Nilansha Dargan[1], Katrin Schulz[1]

[1] Institute for Logic, Language and Computation, University of Amsterdam
[2] Department of Experimental Psychology, Utrecht University

Abstract

We present a dataset on how stereotypically "feminine" or "masculine" different occupations are viewed as well as on how familiar raters were with the occupation. As many measurement tools for gender biases in natural language processing (NLP) systems utilize occupational labels, our participants primarily rated occupations that have been used for such NLP measurement tools. We hope that our dataset will be used to assess the extent to which language model biases correspond to human stereotypes. To that end, we provide access to an application for downloading our data that enables users to customize the dataset, depending on their practical needs and theoretical beliefs.

To test for the robustness of our gender stereotypicality ratings, we collected data across 2 scale directions, 2 different questionnaire designs and 3 different sample makeups (ethnically representative UK and US samples, as well as a less representative US sample). We found that ratings were relatively stable across scale directions and questionnaire design and that raters' gender and age were not predictive of the ratings they gave.

One of the included questionnaire designs is replicated from Shinar (1975) – which had already been partially replicated by White et al. (1989). This allowed us to compare stereotypicality ratings from three different timepoints. We found that ratings from Shinar (1975) were more extreme than ratings from the two studies thereafter.

4.1 Introduction

It is a difficult but consequential task to decide which candidate should be invited to a job interview, and, following such interviews, which candidate should be hired. When making such decisions, humans are flawed, relying on their biased intuitions and overly prioritizing aspects of applicants, like physical appearance, or similarity to the interviewer, that are of little importance in work-life. It would be wonderful if one could simply click a button, and have a perfectly objective computer genie tell one whom to hire, no interview required. In 2014, Amazon apparently set out on a journey towards this "holy grail" (Dastin, 2018) of hiring tools. Their experts on machine learning developed a system to predict, based on applicants' résumés, who should be hired.

Unfortunately, when such machine learning systems detect rules and associations in the data they are trained on, not all of those are helpful or desirable. The Amazon employees noticed that résumés that included the word "women's" – for example, in "women's chess club captain" – were rated more poorly by the system (Dastin, 2018). Apparently, because most of the résumés the system was trained on stemmed from men, it learned that these womanly words are not befitting of an (Amazon) applicant.

In the years since, such computerized technology that operates on language – Natural Language Processing (NLP) technology – has become more wide-spread (see e.g., Gottfried et al., 2026). Few examples of discriminatory NLP behavior are as blatant as the one of the scrapped Amazon system. And yet, also modern NLP technology, like the Large Language Models (LLMs) that underlie much of today's generative language technology (e.g., multi-purpose chatbots like ChatGPT), pick up on rules and associations on their own. Consequently, they, too, can acquire associations that are undesirable and, if acted upon, potentially harmful (e.g., while Amazon employees did not solely rely on outputs of their résumé evaluation system, they did consider them; Dastin, 2018 – meaning that some applicants might have been impacted by it).

Given how widely-applied NLP technology is – including, at times, to support high-stakes decisions like hiring, and potentially (as that use case is being explored; Zack et al., 2024) medical diagnoses – measuring and mitigating biases in NLP systems has emerged as an active domain of NLP research. As we discussed in detail in van der Wal*, Bachmann* et al. (2024), existing measurement tools of such biases have several weaknesses. These include their poor ability to predict problematic behaviors ("downstream harms") when the assessed systems are applied (Goldfarb-Tarrant et al., 2021), as well as the fact that measurements of the same system from these different tools often align poorly, if at all (Asgari et al., 2026; Cao et al., 2022; Delobelle et al., 2022) – two observations which suggest that the measurement tools either assess different underlying properties of the NLP system (or nothing meaningful at all). Existing NLP bias measures will need to be improved – or better ones will need to be developed – and we

previously discussed (in Bachmann et al., 2024; van der Wal*, Bachmann* et al., 2024) our views on how such development and evaluation efforts could be improved.

Many NLP bias measurement tools, especially for gender biases, involve the system acting on occupational terms (e.g., the highly influential WEAT/WEFAT; Caliskan et al., 2017; WinoBias; Zhao et al., 2019; and WinoGender; Rudinger et al., 2018, but also more recent ones like Döll et al., 2024 or OccuGender; Chen et al., 2025). That is because occupational labels are nominally non-gendered but there are many gendered stereotypes about occupations: For example, if a language model is much more likely to predict the next output to be "he went home." rather than "she went home.", given the context "Because the doctor was tired, ", it apparently considers "doctor", a high status occupation, to be a "masculine profession". To test whether such system behaviors correspond to human stereotypes (e.g., whether systems can be claimed to perpetuate human stereotypes), an outside criterion – some "ground truth" about whether an occupation is stereotypically "masculine" or "feminine" – must be considered. NLP researchers often use "percentage of women and men employed in this occupation" as a proxy of how stereotypically "feminine" or "masculine" an occupation is (e.g., Bartl et al., 2020; Mattern et al., 2022; Zhao et al., 2019, i.a.).

As we noted in van der Wal*, Bachmann* et al. (2024), this proxy is imperfect: Not only can there be occupations where employment statistics are directionally misaligned with gender stereotypes – for example, an occupation seen as stereotypically masculine despite more women working in it – there can also be meaningful differences in magnitude, like an occupation seen as "very feminine" despite women only making up a slight demographic majority. We assume that the frequent use of such Labor Statistics data is mostly born of convenience – that data of labor market participation is readily available, but data on gender stereotypical views about occupations is not.[1] After all, if the goal is to prevent harms of NLP technology, we believe that conclusions like "The language models' outputs include (human) stereotypes" more meaningful than "The language models' outputs coincide with labor market participation rates".

With this paper we aim to make accessing relevant data about gender stereotype beliefs convenient. We created a dataset of people's stereotypical views about 210 occupations – many of which were selected because they have in the past been used for NLP gender bias measures (see Section 4.2.1) – and simplified customization of the dataset, depending on users' practical demands and theoretical considerations.

[1] A note-worthy exception to the rule is Farkas and Németh (2022) who collected questionnaire data similar to ours to assess the consistency between gender stereotypes and Google Translate's gendered translations.

Overview of this paper Our primary goal for this paper was to provide NLP researchers with a dataset they can use to validate (i.e., systematically test whether a measurement tool truly captures the underlying property it is meant to quantify; the topic of van der Wal*, Bachmann* et al., 2024) the bias measurement tools they develop. That being said, we had two additional goals and target audiences in mind.

As people's gender stereotypical views about occupations are also informative for social scientists (e.g., organizational psychologists), we took this opportunity to collect data in a format consistent with two earlier social psychology papers that obtained similar data several decades ago: Shinar (1975) and White et al., (1989). This allows us to observe whether gender stereotypical beliefs about occupations differ across decades or are relatively stable (for more direct replication efforts of Shinar's original paper, see Beggs & Doolittle, 1993; Munoz Sastre et al., 2000).

Finally, as design decisions about questionnaires are not trivial and people have differing views on what constitutes best practice, we implemented different versions of the gender stereotypicality questionnaire. Our dataset might hence be interesting to some psychometricians – psychological researchers specialized in the development and evaluation of measurement tools for properties or contents of the human mind (e.g., intelligence or happiness) – that are curious to see (or to themselves test with more sophisticated statistical methods) which design choices made differences in practice.

As we hence have three different audiences for whom some but not all of our sections are relevant, we marked sections with whom they are aimed towards: [N]LPers, [P]sychometricians, and/or [S]ocial psychologists. All "paths" through the paper are meant to be independent: A social psychologist should be able to understand a section marked with [S], while only having read all previous sections marked with [S]. How a section is marked also influences the background knowledge we presume readers to possess (e.g., in sections marked relevant for NLPers, we do not presume that readers are familiar with linear mixed effects models).

4.2 Methods

[N] [P] [S] **Overview** For 210 occupational labels (see "Included Occupations" in Section 4.2.1), we collected participants' ratings of two attributes: the extent to which the occupation is seen as "feminine" or "masculine" and the extent to which the occupation is familiar to them. Both for the gender stereotypicality and the familiarity ratings, occupations were shown one-by-one, and the order in which occupations were shown was randomized. First, participants gave gender stereotypicality ratings for all occupations. We implemented four different versions of these stereotypicality ratings (see Section 4.2.1), one of which each

participant was randomly assigned to.

Subsequently, replicating Shinar (1975), we asked with a multiple-choice question which decision criteria participants had applied, when rating the stereotypicality of occupations. Based on which questionnaire version a participant performed, they saw one of two versions of this criterion question. Both of these versions had identical answer options (see Figure 4.1). If participants answered that they used a different criterion than the ones listed (i.e., chose answer option 4), we subsequently asked them to specify in an open text box. If participants indicated using multiple criteria, we asked them to "provide a rough estimate of how much each of the different criteria [...] weighed in [their] judgment": We listed all criteria they had previously indicated using, and participants gave integer percentage estimates of how much each weighed in their decision (i.e., across criteria, entered values had to add to 100).

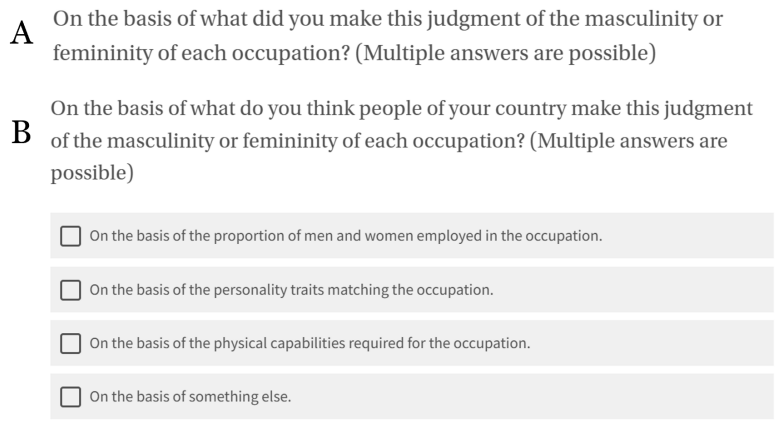


Figure 4.1: Multiple-choice question asked to participants after they gave ratings on the gender stereotypicality of different occupations. Question phrasing, but not response options, differ based on the questionnaire version a participant completed (see Section 4.2.1 for details on these versions): A) Phrasing for the two versions that replicate Shinar (1975) B) Phrasing for the other two versions.

Finally, participants indicated their familiarity with the occupations they had just rated on gender stereotypicality. We used Schröder et al. (2012)'s answer categories and used a slightly adjusted version of their explanation text (categories and explanation text both in turn are based on Snodgrass & Vanderwart, 1980). For both, see Figure 4.2. We collected familiarity ratings for two reasons: firstly as an optional filter to exclude stereotypicality ratings from participants for occupations they do not know. Secondly so that we do not falsely exclude participants, assuming that they are inattentive, while their unusual response to an occupation actually stems from them not knowing the occupation (see Section 4.2.2).

Data collection occurred online as questionnaires were implemented on Qualtrics

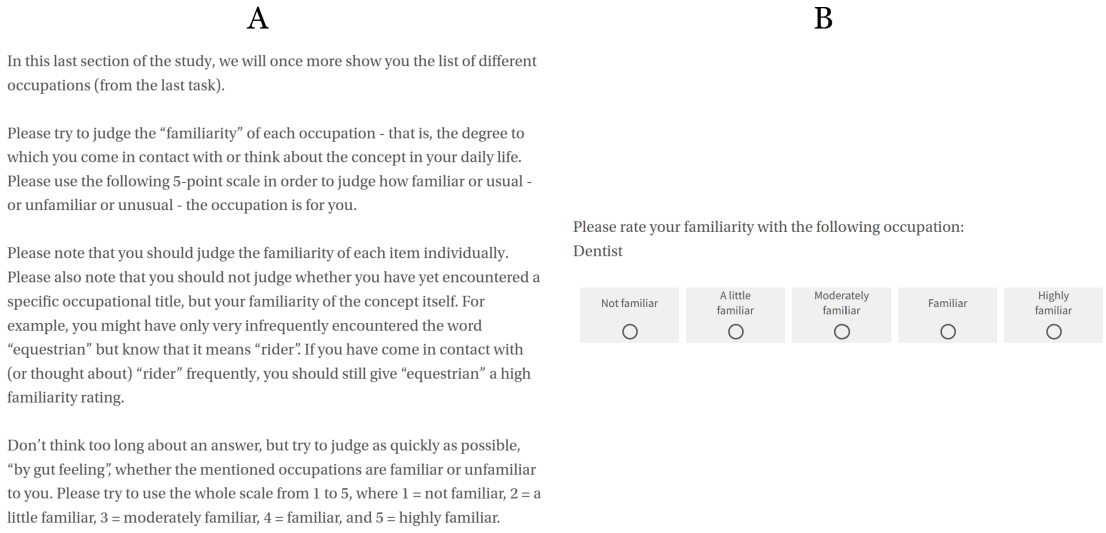


Figure 4.2: A) instructions that were provided to participants before they started rating how familiar they were with the different occupations. B) example of a familiarity rating. Participants in all four questionnaire conditions were exposed to the same familiarity question and scale direction.

(https://www.qualtrics.com/). The median completion time of our study was 32.72 minutes (range: [15.6, 99.1 minutes]. All data and material needed to replicate this study are available at https://osf.io/ep8gn/overview?view_only=d64bac1e707b4aa1a8c9d6bbb50b95f1. We preregistered this study (i.e., determined exclusion criteria and analysis approach before data was collected; on the benefits of preregistration see e.g., Nosek et al., 2018) at https://aspredicted.org/t837-fv2h.pdf. For deviations from our preregistered protocol and why we made them, see Appendix A.1.

4.2.1 Questionnaire design

[P] [S] **Implemented versions of gender stereotypicality part** Participants rated the gender stereotypicality of occupations with one of four questionnaire versions. The first two versions are a direct replication of Shinar (1975)'s original 7-point scale. This scale has verbal labels at the first, fourth, and seventh scale point – "Feminine", "Neutral" and "Masculine" – but not at the remaining 4 scale points (see I.B of Figure 4.3). The second version of the questionnaire had identical answer categories and near-identical instructions[2] but a scale direction of "Masculine" to "Feminine".

Alongside this scale, we also included Shinar's original instructions – insofar

[2]Instructions differed, because they were held consistent with participants' scale direction: Half read "on a scale from masculine to feminine", the other half "[...] from feminine to masculine".

as we could reconstruct them based on what she mentions in her paper (see I.A of Figure 4.3). Shinar (1975) herself compared stereotypicality ratings across three different instructions and White et al., (1989) replicated two of her instructions, but we replicate only one of the three here. For information on the non-included instructions, as well as other changes we made in setup compared to Shinar (1975)'s original study, see Appendix A.2.

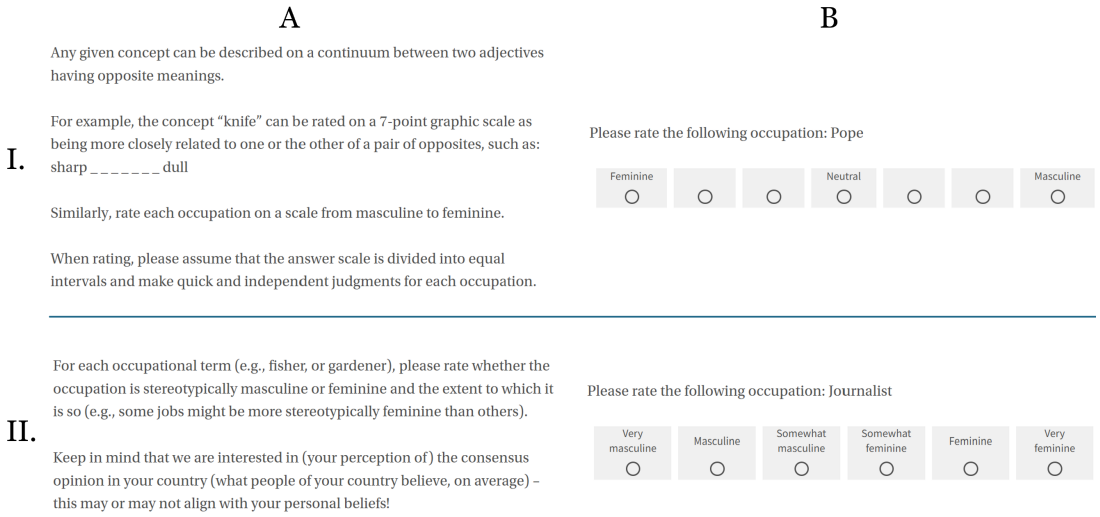


Figure 4.3: A) instructions provided before participants started rating and B) examples of different versions of the gender stereotypicality rating question. I.) pertains to the 7-point scale that replicates Shinar (1975), II.) to the fully-labeled 6-point scale. While we here show a "feminine first" 7-point scale and a "masculine first" 6-point scale, oppositely scaled variants (see e.g., I.A) also existed.

Our two other versions of the stereotypicality ratings included a response scale that we deemed more methodologically sound for our purpose: Instead of a partially labeled 7-point scale, participants in these groups responded on a fully labeled 6-point scale from "Very feminine", to "Very masculine" (see II.B of Figure 4.3; incidentally, those are the response categories also used by Farkas & Németh, 2022). As was the case for the two Shinar versions, these two versions differed in their scale direction and instructions, one going from "Very feminine" to "Very masculine", the other from "Very masculine" to "Very feminine".

We utilized a 6-point scale, as it forces participants to commit to a directionality in their response, instead of them being able to simply respond with the mid point of the scale. Our questionnaire should hence be less vulnerable to socially desirable responding (Grimm, 2010) – participants answering with what they feel they should answer rather than with what they genuinely believe – and more sensitive to subtle stereotypes; those where participants might "round" their response to a neutral midpoint even though they have small inclinations about whether an occupations is feminine or masculine. A potential downside of

using a 6-point scale is that our questionnaire could be less sensitive for detecting occupations that participants genuinely feel neutral about. We are not too concerned about this: If participants, as a whole, see an occupation as neutral in terms of gender stereotypicality, this should still come out on aggregate: We would then expect that on average half of the participants who'd prefer a neutral option choose "somewhat feminine" and the other half "somewhat masculine". Scale direction could matter here – participants could satisfice and choose the first response category they encounter that is close to the neutral option they'd wish to answer (e.g., "somewhat female" for participants whose scale starts with "Very feminine") – but such effects, too, should aggregate out, as both scale directions occur equally often amongst participants.

A further change that we made compared to the 7-point scale is that we implemented a fully labeled scale. We believe that this should make response options less ambiguous: For example, instead of having to distinguish between "1 point from Feminine" and "2 points from Feminine", participants encounter explicit degrees of femininity which they can match with their internal impression of an occupation.[3] A final change compared to the two Shinar replication scales is that we adjusted the instructions (see II.A of Figure 4.3). Besides necessary adjustments in regards to the response categories (e.g., we do not include Shinar's inapplicable mention of "equally spaced" response options), this also involved asking participants to report what they perceive to be the consensus opinion in their country, instead of simply asking participants to rate the occupations. By creating this degree of separation between rater and rating, we hope to decrease the chance that participants respond in socially desirable rather than genuine ways.[4]

[N] [S] **Included occupations** The occupational terms that participants evaluated stemmed either from NLP papers that investigated algorithmic gender bias or from psychology papers that collected ratings of gender stereotypicality for such terms.

Specifically, we first included all terms from the NLP Gender bias challengesets WinoGender (Rudinger et al., 2018; whose list of occupations is based on Caliskan et al., 2017) and WinoBias (Zhao et al., 2018)[5] as well as the 27 terms that White

[3]This comes at a potential cost: Fully labeled response options could be less equidistant than partially labeled ones (Höhne et al., 2021, e.g., there might be a bigger distance in opinion between "Very feminine" and "Feminine" than between "Feminine" and "Somewhat feminine"), meaning that mean aggregating across responses might be misleading.

[4]It is reasonable to object that "perceived gender stereotype consensus" is a different construct from "gender stereotype". We considered the trade-off for fewer socially desirable responding worthwhile, especially given that the perceived consensus opinion – provided it is mostly accurate – is arguably more important than a participant's particular idiosyncratic views on occupations and gender.

[5]Due to a typo (i.e., a missing separator), we did not include WinoBias's "chief": We wrote and hence received ratings for "guard chief", instead. As the resulting occupational label

et al., (1989) used to measure changes over time compared to Shinar (1975).[6]

Subsequently, we included one-word labels[7] from other sources. For words from these sources, we curated: We excluded terms we deemed obscure (e.g., "phlebotomist"), ambiguous in meaning (e.g., "controller"), redundant with already included terms (e.g., "aviator", given the already-included "pilot"), or gendered (e.g., we replaced "policeman" with "officer"). We first included terms from psychological papers that collected gender stereotypicality ratings (i.e., Adachi, 2013; Shinar, 1975; White et al., 1989), then terms from Farkas and Németh (2022) who collected such ratings to address algorithmic gender bias[8], subsequently from StereoSet (Nadeem et al., 2021), an NLP benchmark that addresses stereotypes about occupations (among other stereotype categories), and finally from other papers that included occupational terms to investigate algorithmic gender bias (i.e., Bartl et al., 2020; Kay et al., 2015; Mattern et al., 2022; Prates et al., 2020).

Finally, we included "nun" and "pope" as catch trials. Participants who indicate knowing the terms but do not rate them as feminine and masculine, respectively, probably did not sufficiently pay attention during the study (for details on this and other exclusion criteria, see the "Exclusion criteria" segment of next Section).

4.2.2 [N] [P] [S] Participants

We collected three different samples of about 300 participants, each, on the participant-recruitment platform Prolific (https://www.prolific.com/). Post exclusions (see next section), we had 254 valid cases for a US convenience sample, 248 for a sample paid to be demographically representative of the USA, and 277 for a sample paid to be representative of the UK. This way we could compare responses between two different Western and English-speaking nations – UK representative sample vs US representative sample – as well as responses from samples of the same country but with different demographic makeups: US representative sample vs US convenience sample. To recruit the two "representative"

coincidentally is somewhat meaningful (i.e., ≈ "captain of the guard"), we did not exclude its data from our analyses.

[6]White et al., (1989) adjusted 8 of Shinar (1975)'s 27 terms. Whenever there were differences between the compared terms, we included the term we considered more common (e.g., White et al.'s "flight attendant" as opposed to Shinar's "air steward(ess)") or less ambiguous. See the bottom half of Table 4.2 for our choices.

[7]We prioritized one-word terms because many historically influential NLP bias measures operated on static word representations (e.g., Caliskan et al., 2017), which more straightforwardly represent individual words than multi-word expressions. Consequently, reference ratings for one-word terms are more useful for the evaluation of such tools.

[8]For Farkas and Németh (2022) we could only include the 39 terms they mentioned in their paper, as their data – and hence their list of included occupations – was unavailable, at the time we compiled our list.

samples, we made use of Prolific's paid "representative sampling" feature, which implemented quotas based on self-reported ethnicity, age and sex. The resulting samples were not fully aligned with their intended ethnicity quotas, however (e.g., White participants were over- and other ethnicities underrepresented). For a more detailed discussion on the ethnicities of the samples, see Appendix A.3.

In exchange for monetary reimbursement, Prolific members were permitted to participate, if they were US or UK residents, adult-aged (18+) and indicated speaking English fluently. The final US convenience sample included 119 women, 132 men and 3 participants who responded "other" for gender, with an average age of 41.86 years (sd = 11.89, range: [20, 77]); the US representative sample included 135 women, 110 men and 3 "other"s with an average age of 46.88 (sd = 15.98 range: [19, 94]); and the UK representative sample 144 women, 132 men, and 1 "other" with an average age of 47.31 (sd = 15.99, range: [19, 88]).

[S] **Exclusion criteria** If participants indicated that they were not particularly familiar with an occupational term (i.e., gave one of the two lowest familiarity responses: "not familiar" or "a little familiar"; see Figure 4.2), we did not include their response, when calculating an occupation's mean stereotypicality (and SD) in our rating Table A.4.

Additionally, we excluded participants' data – for all analyses and the table – if they gave responses that are suggestive of inattentive or random responding. Specifically, we included "nun" and "pope" as catch-trials, as they are gendered by definition. Participants that did not, for both, provide the highest or second-highest stereotypicality rating in the correct direction were excluded. This did not happen for participants which gave incorrect ratings for either term but also were unfamiliar with said term (i.e., provided familiarity ratings below "moderately familiar"; see Figure 4.2). We additionally excluded participants, that gave highly differing stereotypicality ratings to very similar terms (i.e., 3 or more rating point differences on "nurse" vs "registered nurse", "mechanic" vs "mechanician", or "programmer" vs "computer programmer").[9]

With these criteria, we excluded 23 participants from the UK representative sample, 49 from the US representative sample, and 47 from the US convenience sample. Additionally, we excluded 4 participants with incomplete data (for more details on these, see Appendix A.1).

[9]Such redundant terms existed, because we had to include some terms verbatim – for example Shinar (1975)'s and White et al., (1989)'s "registered nurse" which our ratings are compared to in Section 4.4.2 – while a more modern and/or one-word term (e.g., "nurse") existed that was worth including.

4.3 The dataset

4.3.1 [N][P][S] Description of the data

Occupations were reasonably familiar, with the average familiarity rating being 3.50 (i.e., a response between "Moderately familiar" and "Familiar"; see Figure 4.2). Only 5 occupations had average ratings below 2.5 (see Figure 4.4). For those, mean values in Table A.4 – especially for the UK – should be interpreted with caution: Since we preregistered that participants ratings only contribute to an occupation's mean rating if they were at least "moderately familiar" with that occupation, few participants contributed to these means.

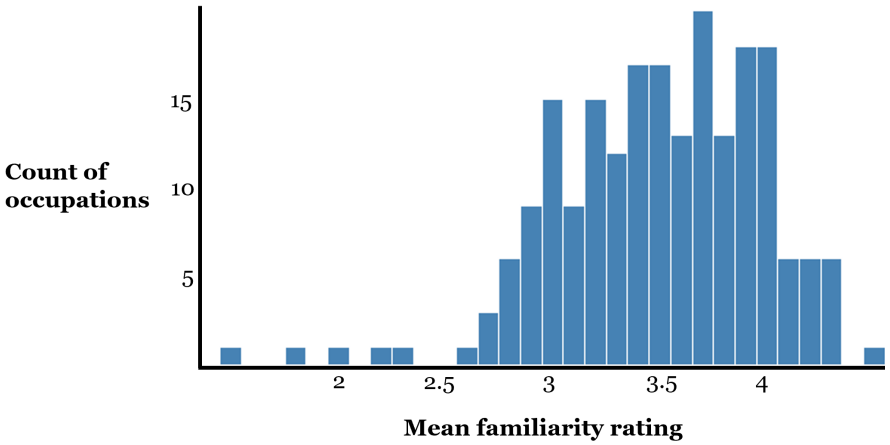


Figure 4.4: Histogram of mean familiarity ratings (x-axis) by count of occupations with such ratings (y-axis). Data is pooled across all included participants. The majority of occupations have mean ratings above "Moderately familiar" (i.e., 3; see Figure 4.2 for all answer categories).

For a breakdown of how often each questionnaire type's response categories were chosen, see Table 4.1. Generally, more "masculine" than "feminine" ratings were provided, both for moderate responses (e.g., more 3s than 5s, and more "somewhat masculine"s than "somewhat feminine"s) and for extreme responses (e.g., more "very masculine"s than "very feminine"s). This tendency was especially pronounced for the 6-point scale (i.e., for whom 67.8% of responses were on the masculine side of the scale), as in the 7-point scale the neutral midpoint received the plurality of responses. These trends were consistent across samples, though UK respondents seemed to generally provide more moderate responses than US respondents (i.e., a higher percentage of midpoint and adjacent-to-midpoint responses for both scales).

These response trends are also visible in Figure 4.5 which displays densities of mean stereotypicality ratings for the different occupations: More means fall to the left than to the right of the neutral midpoints (see dotted lines in the

Table 4.1: Percentage of responses that fell into the different response categories, by scale type (7-point scale vs 6-point scale) and sample.

Shinar (7-point scale)

	Masculine	2	3	Neutral	5	6	Feminine
Overall	11.0%	12.5%	17.0%	40.3%	8.8%	5.8%	4.6%
UK-rep	7.9%	11.9%	17.5%	44.4%	9.2%	5.7%	3.3%
US-rep	10.9%	13.4%	17.5%	39.1%	8.6%	6.0%	4.6%
US-conv	14.4%	12.4%	16.0%	37.0%	8.6%	5.8%	5.8%

Social Desirability Bias version (6-point scale)

	Very masculine	Masculine	Somewhat m.	Somewhat f.	Feminine	Very feminine
Overall	10.2%	26.2%	31.3%	15.9%	11.3%	5.1%
UK-rep	9.1%	23.3%	36.0%	17.3%	9.9%	4.3%
US-rep	10.1%	27.6%	29.7%	15.5%	11.9%	5.3%
US-conv	11.7%	28.0%	27.7%	14.7%	12.2%	5.7%

Figure), and mean responses on the 7-point scale were more moderate (i.e., closer to the neutral midpoint) than responses on the 6-point scale. That the plurality of responses on the 7-point scale are "neutral"/4 responses probably contributes to mean ratings being more moderate for those scales.[10] Scale direction, on the other hand, apparently had little impact on responses, as the yellow densities of the feminine-first respondents are very similar to the blue masculine-first ones.

Responses were similar in kind across the 6 and 7-point scales: Within samples (with responses pooled across scale directions), mean ratings of occupations were consistent across scale type in whether they fell on the masculine or feminine half of their scales: They respectively agreed on 198 (UK), 198 (US representative), and 204 (US convenience sample) out of 210 occupations. Secondly, mean ratings of occupations ranked near-identically across scale types, Kendall's $\tau_{UK} = .924$, confidence interval based on 5000 bootstraps: [.907, .939], $\tau_{USrep} = .915$, CI: [.896, .931], $\tau_{USconv} = .929$, CI:[.914, .942].[11]

[S] **Participants' criteria for judging gender stereotypicality** When rating the gender stereotypicality of different occupations, Shinar (1975)'s partici-

[10] We also inspected a density plot of mean responses from 7-point scale respondents, if neutral responses are deleted and the remaining 6 response options are rescaled to 1-6. The resulting distribution was similarly right-skewed to the distribution of the 6-point scale, $Median_{Recoded} = 2.71$, $Median_{6-Point} = 2.87$, $IQR_{Recoded} = 1.49$, $IQR_{6-Point} = 1.30$. Evidently, 7-point respondents were not per se shier than their 6-point peers to provide extreme responses – provided that an occupation was sufficiently "non-neutral" to receive a response other than the (apparent default) midpoint response.

[11] For readers more familiar with Spearman's rho: $\rho_{UK} = .990$, bootstrapped confidence interval: [.983, .993], $\rho_{USrep} = .988$, CI: [.981, .992], $\rho_{USconv} = .992$, CI:[.987, .994].

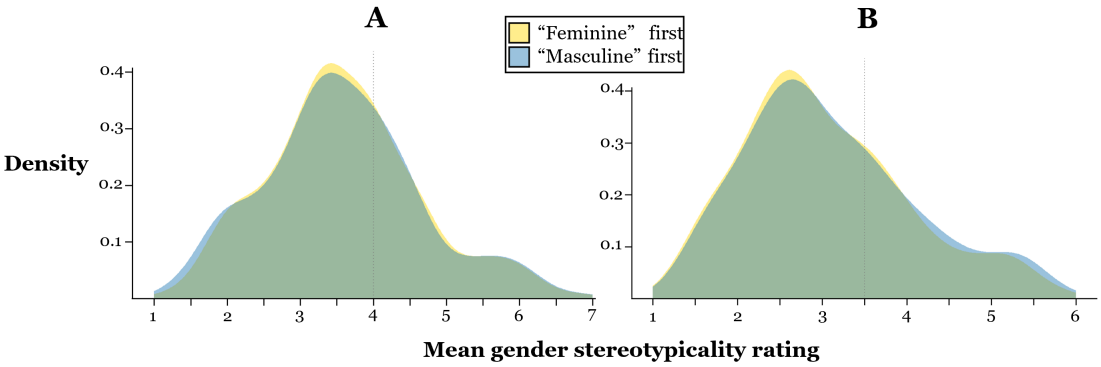


Figure 4.5: Kernel densities of the mean stereotypicality ratings for participants with the 7-point (A) and 6-point scales (B). Dotted vertical lines correspond to the neutral midpoints of the scales. Yellow [blue] densities stem from ratings of participants whose scale started with feminine [masculine]; green shows their overlap. Lower values on the x-axes correspond to more masculine responses.

pants predominantly reported using the "proportion of men and women employed in the occupation" (i.e., 86% reported using this criterion with 73% reporting that it was their primary criterion) and "the personality traits matching the occupation" (i.e., 67% used it, 27% primarily). Only 3 and 1% of participants, respectively, reported using the other two criteria: "physical capabilities required for the occupation", and the catch-all "other".

In our sample, the three most common response patterns were to pick all criteria except for (our version of) "other", 33.5%, to choose the proportion criterion and the physical capabilities criterion, 21.7%, and to solely choose the proportion criterion, 12.1%. Overall, as was the case for Shinar, the perceived gender proportion was the most important criterion: It was the most commonly chosen one, with 80.9% of participants selecting it as one of their criteria. It was also most frequently the primary criterion: When participants reported employing it as one of multiple criteria, it received in 65.9% of cases the largest or tied-for-largest share of subjective weight (see Section 4.2) in determining participants' stereotypicality ratings.

In contrast to Shinar, the "physical capabilities required for the occupation" was commonly picked: 74.5% of participants chose it as one of their criteria, compared to 56.4% for the personality criterion, and 10.8% for the catch-all "other" criterion. The latter "other" criterion seems to address idiosyncratic criteria that are relatively important to participants: This criterion received the highest or tied-for-highest proportion of subjective weight in 37.9% of cases in which it was among the chosen criteria, compared to physical capability's 35.8 and personality's 31.1% of cases in which those were among the chosen criteria. Additionally – always second to the "proportion" criterion – it received the second-highest mean percentage scores for participants that chose two criteria, $M_{2Proportion} =$

59.6%, $M_{2Other} = 50.9\%$, $M_{2Physical} = 43.0\%$, $M_{2Personality} = 45.7\%$, three criteria, $M_{3Proportion} = 40.3\%$, $M_{3Other} = 32.3\%$, $M_{3Physical} = 31.0\%$, $M_{3Personality} = 28.7\%$ or all four criteria, $M_{4Proportion} = 29.0\%$, $M_{4Other} = 26.6\%$, $M_{4Physical} = 25.6\%$, $M_{4Personality} = 18.8\%$.[12]

There were no noteworthy differences in response patterns between the different samples and questionnaire types, with one exception: We have some evidence that 6-point scale raters ascribed the "Physical characteristics" criterion more frequently to others than 7-point scale raters reported using the criterion, themselves (for differences in instructions between these groups, see Figure 4.1A vs B), 79.0% vs 69.6%, BF = 5.8. For details on this exploratory comparison, see Appendix A.4.

[P] Neutral responses In total, 4 participants in the Shinar replication versions responded with the 4th option "neutral" on all non-catch trials, apparently expressing the view that no occupation is masculine or feminine besides occupations gendered by definition. Three further participants provided this response in more than 95% of non-catch trials and one further participant was excluded from the study, because she provided this neutral response on all non-"pope" trials, including the "nun" catch trial.

For the questionnaire versions with 6 response categories, the closest equivalents to always choosing a neutral 4 option is to always choose the 3rd and/or the 4th option "somewhat masculine" and "somewhat feminine". None of our participants did so, though 5 participants provided such responses in 95% or more of non-catch trials. Additionally, two participants were excluded for providing such responses on catch trials, as they respectively providing such responses on all and on all-but-two trials.

We were curious about whether the 5 non-excluded participants that chose the 3rd or 4th option near-exclusively (i.e., on $\geq 95\%$ of trials) chose between them randomly – for example, because they have fully neutral beliefs but cannot express them on our scale – or whether their responses still reflected stereotypical beliefs. As an exploratory analysis, we hence compared two hierarchical logistic regression models, one of which included, as a predictor, standardized mean ratings of occupations – from all bar the 5 assessed participants. As described in more detail in Appendix A.4, other participants' standardized mean responses were highly predictive of whether these 5 participants responded with a 4 or a 3, and the model that included others' means as a predictor was favored decisively, $BF_{ModelComparison} \approx 10^{54}$. Participants that near-exclusively answered with 3s and 4s evidently still expressed stereotypical beliefs therewith.

[12] While we are moderately confident that this "other" option captures relevant idiosyncratic criteria, its apparent status as "the second most important criterion, if included" is less certain: Since it was rarely chosen in absolute terms, its percentage values are based on much fewer observations than those of the other criteria (i.e., based on only 84 participants, compared to the ≥ 439 participants of each other criterion).

4.3.2 [N] How to access the OccupatiEN data

The gender stereotypicality means and standard deviations in Table A.4 stem from ratings that were pooled across different questionnaire versions (i.e., those that we found to be about equivalent in our statistical analyses in Section 4.4.1). That being said, your particular task demands (e.g., if it addresses a particular age subgroup) and differing theoretical preferences – for example, in regards to question phrasing (see Section 4.2.1), or participant exclusion criteria (see Section 4.2.2) – may make you wish we had made different choices in regards to how we compiled the table. We previously stressed the importance of such pluralism – provided one's choices are explicitly communicated (van der Wal*, Bachmann*, et al., 2024). To hence facilitate the flexible use of our dataset, we created a R Shiny app whose source code is accessible, alongside our anonymized raw data at https://osf.io/ep8gn/overview?view_only=d64bac1e707b4aa1a8c9d6bbb50b95f1.

R Shiny apps are interactive applications which allow users to execute R code on the backend, by manipulating simple (e.g., point-and-click) user interfaces (Chang et al., 2026). In the left panel of our user interface, depicted in Figure A.1, you indicate your customization preferences. Given these choices, 3 outputs are dynamically updated in the central panel of the application window: A text describing the choices you have made (to simplify communicating that information; e.g. in a paper), a code that succinctly labels these choices (read the caption of Figure A.1 for more details) and a preview of the table of stereotypicality ratings resulting from these choices. If the previewed table looks acceptable, its full version can be downloaded as a .csv file.

A word of caution for creators of NLP tools (e.g., gender bias challengesets) who wish to use our ratings to validate their tool: Do not "shop" for strong alignments between the tool and our ratings by trying out different versions of our dataset. Instead pick a singular version of this dataset that you find most applicable to your aims (theoretically and practically), else you are likely left with an overly optimistic view on your tool.

4.4 Inferential Analyses

4.4.1 [N] [P] [S] Which factors impacted stereotypicality ratings?

To determine which factors of the participants and the questionnaire design influenced the obtained stereotypicality ratings, we ran two parallel series of exploratory analyses, one each for participants that responded to scales with 6 versus to scales with 7 response options. Each time we aimed to assess whether the following four factors were predictive of how participants rated occupations: Their age, gender, sample membership (e.g., whether UK raters generally gave

less masculine responses than representative US raters) and the direction of the response scale they encountered (i.e., whether their scale started or ended with the feminine half).

We assessed the importance of the four factors by fitting models that predicted standardized stereotypicality ratings based on one of the 16 possible combinations of the four factors (e.g., a model that included all factors, 4 different models that miss a factor but include the three remaining ones, etc.) as well as based on random effects predictors which were identical between all models (for details on these, see Appendix A.4). After fitting all models, we could evaluate the importance of an individual factor by comparing how well the 8 models that included it account for the data compared to the 8 models that did not include it.

Before fitting models we standardized the to-be-predicted stereotypicality ratings, separately for all participants that had 6 and for all participants that had 7 response options. Predicting standardized rather than raw stereotypicality ratings allowed us to compare parameter estimates across our two parallel series of model comparisons (and to use the same prior distributions for both model series; see Appendix A.4).

Unlike what we initially planned, we ended up pooling data between our two US samples, only contrasting US participants with UK participants instead of contrasting all three samples. For the first set of models that we fit, random slopes between the US representative and US convenience samples were near perfectly correlated, $\rho_{6:USc\&USr} = .99$, 95% equal-tailed credibility interval: [.97, 1], $\rho_{7:USc\&USr} = .97$, CI: [.94, .99]. Such a perfectly collinear random-effects structure made bridge sampling – a prerequisite for our way of comparing different models – impossible, necessitating the pooling. Since the two samples also had very similar estimates for all their parameters (e.g., for the 6-point scale model: $M_{USc-UK} = -0.03$, CI: [-0.09 , 0.02] vs $M_{USr-UK} = -0.01$, CI: [-0.07 , 0.04]); $\tau_{USc} = .21$, CI: [.19, .23], $\tau_{USr} = .19$, CI: [.17, .22], $\rho_{UK\&USc} = .22$, CI: [.08 , .36], $\rho_{UK\&USr} = .20$, CI: [.06 , .33]), that is probably not a big loss: There seem to be virtually no differences between the ratings of these two samples.

When evaluating the different factors, the substantive takeaways for the two analysis series were similar: The inclusion Bayes factors provided moderate or strong evidence that a rater's gender, $BF_{6PointScale:ExclusionOfGender} = 3.04$, $BF_{7:excl} = 11.75$, age, $BF_{6:excl} = 430.70$, $BF_{7:excl} = 233.61$ and scale direction, $BF_{6:excl} = 3.08$, $BF_{7:excl} = 19.59$ were not predictive of the ratings they gave. Additionally, the occupation-based random slopes of both model series suggested that there were modest differences between UK and US participants in how feminine or masculine they rated different occupations; for all 6-point scale models: $\sigma_{UKvsUS} = .20$, CI: [.18, .23], for all 7-point scale models: $\sigma_{UKvsUS} = .20$, CI: [.18, .22].[13]

[13]For an intuition about the size of these differences: The (model-implied) occupation-based

The only substantive difference in results is that we obtained strong evidence against the relevance of the UK vs US contrast main effect for the 6-point scale, $BF_{excl} = 14.83$, while we obtained weak evidence for its relevance for the 7-point scale, $BF_{inclusion} = 2.47$. Consistently across 7-point scale models, US participants provided slightly more masculine responses than UK participants (i.e., for 6 of the 8 models, the estimates and credible interval were $\beta_{UKvsUS} = -0.07$ CI: [-0.12, -0.02], for the remaining 2 models, the upper bound of the CI was -0.01). We presume that this difference between countries stems from the 7-point scale's neutral midpoint option: Compared to the other two samples, UK respondents provided more such neutral responses and provided fewer extreme responses on either side of the scale (see Table 4.1). As there were generally more responses on the masculine than feminine half of the scale for all samples, both trends – fewer extreme responses and more midpoint responses – increase the mean response rating (i.e., make the UK sample's average response "less masculine").

As is the case for all other Bayesian analyses reported in this paper, information on the convergence of and prior distributions for these analyses can be found in Appendix A.4.

4.4.2 [S] Comparisons with previous gender stereotypicality ratings

To test whether ratings have become more gender neutral across the three testing instances, we performed a Bayesian analysis of variance with a "studies" factor (Shinar vs White et al. vs our study) on 26 extremity scores for the ratings of terms that are similar between White et al., (1989) and Shinar (1975) – that is: all terms in Table 4.2 except the "educational administrator" / "school administrator" pair. By "extremity score" we mean the number of scale points that separates the mean response from the neutral response 4 (e.g., means of 1 and 7 would both have extremity scores of 3).

differences between UK and US ratings are approximately normally distributed with a mean of 0 and a standard deviation of 0.2. That means that for about 68% of occupations, the UK–US difference is expected to lie within ±0.2 standard deviations (small differences), and for about 95% within ±0.4 standard deviations (small to moderate differences) of the stereotypicality ratings.

Table 4.2: Means and (in brackets) standard deviations of gender stereotypicality ratings from our US convenience sample with the 7-point response scale – as well as White et al. (1989)'s, Shinar (1975)'s and Beggs and Doolittle (1993)'s ratings with the same scale – on 27 shared occupations. While White et al. have differing numbers (see N_{White}), each of our means is based on 127, each of Shinar's on 120, and each of Beggs and Doolittle's on 142 participants. The ratings below the line are based on similar but non-identical occupation labels; in italics we indicate which of the studies used a different label from ours, and which label they used. In keeping with tables from these studies, occupations are sorted from "masculine" to "feminine" (i.e., as per White et al.'s ratings). No standard deviations are listed for Beggs and Doolittle, as they did not report these. Besides these occupations, we also share 29 occupations with some but not all of the previous studies (see Table A.3).

Occupation	Bachmann	White	N_{White}	Shinar	Beggs
Carpenter	1.80 (0.91)	1.90 (1.07)	174	1.75 (0.39)	1.60
Electrician	1.95 (1.00)	1.99 (0.96)	177	1.58 (0.45)	2.30
Engineer	2.42 (1.11)	2.35 (1.00)	176	1.92 (0.45)	2.67
Architect	2.75 (1.08)	2.73 (0.97)	177	2.50 (0.82)	2.67
Dentist	3.31 (1.16)	2.73 (1.08)	177	2.08 (0.62)	2.66
Physicist	2.95 (1.19)	3.01 (1.10)	177	2.25 (1.11)	2.75
Geologist	3.13 (1.17)	3.16 (1.05)	177	2.75 (1.29)	2.83
Physician	3.32 (1.13)	3.26 (0.95)	177	2.67 (1.33)	3.40
Veterinarian	4.21 (1.26)	3.28 (1.04)	177	2.67 (1.51)	3.53
Pharmacist	3.85 (1.25)	3.30 (1.00)	175	3.00 (1.27)	3.63
Mathematician	2.77 (1.17)	3.37 (0.95)	177	3.17 (1.79)	2.97
Computer Programmer	2.61 (1.05)	3.44 (0.87)	177	3.42 (0.63)	3.69
Public Relations Director	4.34 (1.28)	4.05 (0.93)	177	3.33 (0.97)	3.95
Personnel Director	3.94 (1.10)	4.06 (0.81)	176	3.50 (1.00)	3.69
Occupational Therapist	4.57 (1.30)	4.15 (0.77)	177	4.83 (1.42)	4.11
Dental Hygienist	5.28 (1.34)	5.15 (1.44)	176	5.83 (1.24)	5.70
Social Worker	5.42 (1.30)	5.22 (0.99)	177	4.75 (0.75)	5.56
Dietitian	5.16 (1.17)	5.31 (1.11)	177	5.25 (0.93)	5.81
Registered Nurse	6.13 (0.95)	5.99 (1.02)	176	6.58 (0.26)	6.00

Farm Manager *White: Agribusiness Manager*	2.33 (1.12)	2.92 (0.97)	177	1.92 (0.99)	2.67
School Administrator *Shinar: Educational Administrator* *Beggs: Dean*	4.32 (1.28)	3.23 (0.86)	177	3.25 (0.75)	3.30
Life Insurance Agent *Shinar/Beggs: Insurance Agent*	3.52 (1.04)	3.31 (0.95)	177	2.83 (1.24)	3.06
Journalist *White: Reporter*	4.11 (0.86)	3.88 (0.71)	177	3.83 (0.15)	4.11
X-Ray Technician *White: Radiologic Technologist*	3.90 (1.18)	4.02 (1.02)	177	4.75 (1.11)	4.26
Realtor *Shinar: Real Estate Sales Work* *Beggs: Real Estate Sales*	4.37 (1.29)	4.15 (0.86)	177	3.42 (1.17)	4.05
Elementary School Teacher *White: Elementary Teacher*	5.90 (1.07)	5.46 (1.08)	177	5.58 (0.45)	5.69
Flight Attendant *Shinar: Air Steward(ess)* *Beggs: Steward(ess)*	5.79 (1.33)	5.86 (1.16)	176	5.75 (3.45)	5.84

As preregistered, we included data from our US convenience sample for participants that responded to Shinar's 7-point response scale (see Section 4.2.1), as we a priori considered this sample most comparable to the samples of the previous studies. As a Bayesian t-test provided moderate evidence that scale direction did not impact ratings for these occupations, $BF_{01} = 3.59$, we included data from participants of both scale directions – as Shinar did as well.

Results indicated that there is too little information to determine whether or not there are differences in extremity between the three studies, $BF_{01} = 1.40$. As preregistered, we repeated this analysis pipeline for only those 19 occupations that are identical across all 3 studies (i.e., the occupations above the line in Table 4.2), to test for the robustness of our conclusion. Result were identical in interpretation, $BF_{01} = 1.68$.

As an exploratory analysis, we fit Bayesian linear mixed models with random intercepts for occupation. Since all three studies included extremity scores for the same 26 occupations, such a random intercept accounts for differences between occupations which could overshadow differences between the three studies (e.g., accounts for the fact that "carpenter" generally will have less neutral mean scores than "journalist"). A model that included a fixed "studies" predictor and a random "occupation" intercept explained the data substantially better than a model that only included the random intercept, $BF_{10} = 1015.70$, indicating that studies differed in extremity. This difference seems to stem largely from Shinar's ratings being more extreme, with a Median difference compared to our

scores of +0.30, 95% equal-tailed credibility interval: [0.15, 0.46]. Unlike what we predicted, our scores were not less extreme than White et al. (1989)'s: Differences were small and trended towards White et al. having less extreme scores, $\triangle Median_{Bachmann-White} = -0.08$, CI: [-0.23, 0.07]. For more details on these analyses, see Appendix A.4.

4.4.3 [N] [S] Consistency of ratings and occupation statistics

To quantify the extent to which objective occupation statistics are consistent with the ratings we obtained (e.g., whether amongst our included occupations, the most-feminine rated one also had the highest percentage of women working in it), we performed a Bayesian Pearson correlation test on rank-transformed versions of these variables. This is a Bayesian analogue to frequentist Spearman rho correlation tests.

Specifically, we correlated the percentage of women that worked in a particular occupation in 2023 – the year in which our participants provided ratings – with standardized stereotypicality ratings from our samples. We standardized so that we could pool ratings across our 6- and 7-point scales, as we judged them sufficiently similar to pool based on our earlier analyses (i.e., Section 4.4.1) .

For the US, we found governmental data (U.S. Bureau of Labor Statistics, 2024) on how many women and men worked in a particular occupation for 89, while for the UK, we found equivalent data (Office for National Statistics, 2023) for 97 of our 208 (non-catch trial) occupations. Whenever multiple of our occupations fell under the same label in the governmental data, we retained only our occupation that most closely corresponded to their label (e.g., for the US, we excluded "programmer" and retained "computer programmer', given their label "computer programmer"). If one of their labels included multiple occupations for which we obtained separate ratings (e.g., our "musician" and "singer" both fit the "musicians and singers" label), we excluded these occupations. If our occupation corresponded to multiple labels, we added the totals for these labels (e.g., for our occupation "engineer" we added data for the separate US labels of "civil engineers" and "mechanical engineers"). In total, we could ultimately correlate ratings for 66 occupations for the US data and for 77 occupations for the UK data.

The percentage of women that were employed in these occupations in 2023 were highly, but imperfectly, correlated with how "feminine" they were rated to be; posterior mean of $\rho_{US} = .89$, 95% equal-tailed credible interval: [.84, .94], $M\rho_{UK} = .85, CI : [.77, .90]$. As an exploratory analysis, we also computed the Pearson correlation between the raw rather than the ranked scores. This aligns with a linear and interval-scale interpretation of employment percentages – for example, to treat an occupation with 90% women as "10% more feminine"

than one with 80% women – which is often implicitly (and improperly) adopted when percentages are used as a proxy for gender stereotypicality. Results were similar to those of the correlation on ranked data, $M\rho_{US} = .88, CI : [.81, .93]$, $M\rho_{UK} = .86, CI : [.79, .91]$. For more details on these analyses, see Appendix A.4.

4.4.4 [P] [S] Information provided by midpoint responses

Given that the "Neutral" midpoint apparently was the default response for our 7-point scale respondents (i.e., it received by far the most responses in all our samples; see Table 4.1), there is one additional way in which the extremity of gender stereotypes could be operationalized: Instead of considering occupations' means, one could consider their percentage of responses that are (not) the neutral midpoint. This gives us an indication of what percentage of respondents consider an occupation "sufficiently stereotypical to comment on" (i.e., to provide a non-default/non-neutral response). To quantify the extent to which such a "threshold" operationalization provides conclusions consistent with our mean ratings, we performed exploratory Bayesian Pearson correlation tests between each occupation's percentage of 7-point responses that are neutral, and the extremity of their mean score on responses from the 6-point scale (i.e., the absolute distance between its mean and 3.5, which reflects a gender-neutral value). Both of these indices of how extremely an occupation is gender-stereotyped were consistent with each other, as their values correlated highly; posterior mean of $\rho_{US} = -.89$, 95% equal-tailed credible interval: $[-.91, -.86]$, $M\rho_{UK} = -.87, CI : [-.90, -.83]$. Apparently, conclusions about the extremity of our gender stereotype ratings are robust across such alternative operationalizations of "extremity".

Besides indexing the extremity of gender stereotypes, midpoint responses could also, at least to some extent, be confounded with participants' familiarity with occupations. After all, if participants do not know what an occupation is, they are unlikely to have gender stereotypes about that occupation (e.g., of all occupations, "dramaturg" both had the highest percentage of midpoint responses, 88.6%, and the lowest mean familiarity rating, 1.47, in our UK sample). To quantify the extent to which midpoint responses reflect (un)familiarity, we performed exploratory Bayesian Pearson correlation tests between each occupation's mean familiarity and their percentage of midpoint responses (i.e., in both cases only with data from 7-point respondents). Our data were ambiguous in regards to whether such a confound is a concern, as we respectively received anecdotal statistical evidence against, $BF_{01}^{US} = 2.68$ and for, $BF_{10}^{UK} = 2.05$, such a correlation existing for our two nationalities; posterior mean of $\rho_{US} = -.12$, 95% equal-tailed credible interval: [-.25, .02], $M\rho_{UK} = -.17, CI : [-.30, -.04]$.

4.5 [N][P][S] Discussion

Given the increasing prevalence of natural language processing (NLP) technology in multiple facets of daily and professional life, there has been growing research interest in ensuring that this technology does not act in undesirable and harmful ways. Terms like "harmful" and "undesirable" betray that this is not purely technical work (Schulz et al., 2025): whether outputs are harmful and/or undesirable is highly subjective. Since a lot of people are impacted by NLP technology – at times, without their knowledge or consent (e.g., when their data is automatically evaluated in hiring decisions; Dastin, 2018) – the opinions of a lot of different potentially-affected people should be taken into consideration when making such judgment calls (similar arguments have been made by Blodgett et al., 2020, i.a.).

NLPers do not need to reinvent the wheel, here: Social scientists – and especially the measurement-specialized psychometricians – are experts in assessing people's opinions. Here, we did social scientific research towards NLP research ends: Given that many NLP gender bias measures utilize occupational terms, we believe that our dataset of gender stereotype ratings will be useful either for the creation and evaluation of measurement tools, or for the empirical evaluation of a theoretically important question involving such measures (more on that in a dedicated paragraph, below). We strived to create a high-quality dataset. We collected large samples of responders, paid additional fees to make our samples more demographically representative of their countries, filtered out poor responders (e.g., with catch trials), collected data on how familiar raters were with what they rated, and implemented 4 different versions of how ratings were elicited.

The ratings for our occupations were highly consistent: Our statistical tests provided evidence that ratings did not differ based on participants' age, gender, or the order in which response options were presented to them. Ratings were also consistent irrespective of whether we elicited them with 7-point scales and asked about participants' opinions, or with 6-point scales and asked them about the opinions of people in their country (i.e., ratings of both scale types received inclusion Bayes Factors with similar substantive takeaways, and mean ratings correlated near-perfectly; all point estimates for our 3 different samples: Kendall's $\tau \geq .92$, Pearson's $r \geq .98$).

This is consistent with previous studies which found similar gender stereotypicality ratings across instructions (i.e., Munoz Sastre et al., 2000; Shinar et al., 1975; White et al., 1989, for an overview, see Appendix A.2) – though our questionnaire versions differed more than did their instructions and we had substantially more statistical power (i.e., more sophisticated statistical models and more participants) to detect potential differences. It seems to us – though we did not collect empirical support for that notion, here – that people have quite clear gender stereotypical associations for professions, which makes such ratings robust.

That does not mean that choices in questionnaire design and question wording

are trivial, in general. To the contrary! Given the many non-arbitrary design decisions that come with assessing humans, we would recommend NLP researchers to seek collaborators from disciplines specialized in collecting human opinions, whenever people's opinions are relevant to the research. Elsewhere (i.e., in Bachmann et al., 2024, and van der Wal*, Bachmann* et al., 2024), we additionally argued that such collaborations can be useful, because many disciplines have long wrestled with questions – like "how does one best assess the internal attributes of black boxes?" (e.g., with humans being the black boxes for psychometricians) – that are now relevant to the assessment of LLMs.

[N] When (not) to compare gender stereotype ratings with occupation-based NLP bias measurements Many NLP researchers (e.g., Caliskan et al., 2017) assume that language models' internal representations align with human gender stereotypes or that discriminatory downstream behaviors of language models are consistent with how humans would discriminate in light of gender stereotypes: For example, similarly to how humans with strong gender stereotypes about software developers might be less inclined to hire female candidates, language models might discriminate against female candidates when asked to weigh in on software developer hiring decisions (though cf. Wang et al., 2024, who found a bias against male applicants across several different sectors including those, like construction, that are more stereotypically male). Such consistency between language model behavior and human stereotypes has been hypothesized, among other reasons, to stem from language models being trained on large quantities of human texts (see e.g., Bender et al., 2021; Bolukbasi et al., 2016), in which gender stereotypes are probably implicitly present.

For researchers with such beliefs, the here-presented OccupatiEN dataset can serve as a tool for *validation* (i.e., roughly: the process of evaluating the extent to which your tool measures the not-directly-observable concept you intend to measure; van der Wal*, Bachmann*, et al., 2024): If you assume that LLM's internal representations or harmful downstream behaviors align with human gender stereotypes, and assume that you designed a tool that can measure the relevant internal representations or harmful behaviors stemming therefrom, a critical test for your tool is whether its measurements align with human stereotypes from our dataset.

While we believe that such alignment between human stereotypes and model characteristics is likely, in absence of a gold standard measurement tool for model representations and/or downstream behavior in actual application (see van der Wal*, Bachmann*, et al., 2024), we remain agnostic about whether it exists. If models in reality are meaningfully biased but in a way that is inconsistent with human stereotypes (e.g., if an LLM treats every occupation as masculine), revising tools so that their measurements correlate highly with human gender stereotypes would be contra-productive (i.e., make them less valid).

In our view this leaves two sensible paths forward: If one holds the strong theoretical assumption that such alignment exists, this assumption should be communicated and data on human stereotypes (e.g., our dataset) should be used for validation. If one is agnostic about whether or not such alignment exists, the NLP gender bias tool should first be validated independently of human stereotype data (e.g., tested for how reliably it predicts downstream harms in actual, non-controlled user-LLM interactions). In a subsequent step, the (under such beliefs: empirical) question of whether model behaviors align with human gender stereotypes may then be assessed.

[S] **Our study in the historical context of occupational gender rating studies** Our exploratory analyses provided statistical evidence that our ratings were overall less extreme than Shinar (1975)'s – as White et al., (1989) and Beggs and Doolittle (1993) had concluded to be the case for their ratings (so did Munoz Sastre et al., 2000 – but their ratings stemmed from Spain and France, not the US). Unlike previous studies that compared their ratings with Shinar's, we did not perform a series of (e.g., 129; Beggs & Doolittle, 1993) t-tests and then qualitatively explored how means differed (e.g., whether they were less or more extreme). Instead we tried to find direct statistical support that ratings became less extreme, by transforming means into scores that indicate their extremity (i.e., how far away from the midpoint they were) and comparing these across studies. This approach came at the cost of statistical power, as t-tests can take into account the spread of ratings (through standard deviations) – information that we lost when computing extremity scores. We believe that this is the primary reason for why our preregistered ANOVAs were uninformative, $BF \approx 1$. Once we took into account the random effects structure, however, we did find decisive evidence that our and White et al.'s ratings are less extreme than Shinar's.

Our ratings were not less extreme than the ones from White et al.; if anything, differences trended in the opposite direction. It is unclear what the takeaway is: Shinar's ratings might be outliers, considering that all studies that came after hers have less extreme ratings (i.e., either her sample was an outlier for the time, or opinions were more extreme at the time and subsequently stabilized at a lower level of extremity). It could also be the case that people's opinions have actually become less extreme since White et al. (1989) but that our samples are simply too different for that difference to become evident (i.e., their on average 20-year-old university student sample might be much more egalitarian than our educationally more varied and on average twice as old participants). Causal attributions for other cross-study differences – for example our clear statistical evidence that gender did not uniformly impact stereotypicality ratings, while Beggs and Doolittle (1993) found evidence for female students giving less masculine responses – are similarly ambiguous. The primary goal of our study was to collect opinions that were representative of all adults, which decreases the extent to which

we can meaningfully compare across studies (i.e., across samples, we only have 65 participants aged 24 or younger). Additionally, to decrease the model space (as any additional predictor doubles the number of models that have to be fitted), we did not assess interaction effects. It is possible that those would reveal effects consistent with previous studies on university aged participants, like an age-gender interaction effect of "among women, younger respondents give less masculine responses".

Unlike previous ratings studies – which either sampled occupations systematically across theoretically interesting dimensions (Shinar et al., 1975; White et al., 1989) or elected to directly replicate Shinar (1975)'s selection of occupations (Beggs & Doolittle, 1993; Munoz Sastre et al., 2000) – our sampling of occupations was driven by our intended application: We selected occupations primarily based on whether they can be used in NLP bias measures (i.e., either because they have been used in these before, or because they were one-word terms), and secondarily based on redundancy with similar occupations. This also applied to the occupations we reused from these previous occupational psychology studies (in Table A.3). That being said, we have data for a lot of occupations and from a lot of participants. Consequently, we hope that theoretically meaningful selections of occupations (e.g., based on agency- or communion-related occupational responsibilities; Abele et al., 2008, 2016) will be made to explore how theoretical dimensions relate to our gender stereotypicality or familiarity ratings.

We similarly leave to future work explorations of whether or how participants' self-reported criteria for making gender stereotypicality judgments relate to said judgments, or their demographics. That only 3% of Shinar's but 74.5% of our participants reported using the physical requirements of a job to make such judgments is a striking contrast. In retrospect, we believe it would have been better to replicate Shinar's setup less closely and start the response options of the criterion follow-up question (Figure 4.1) with "On the basis of what felt correct". It is possible that the question in its current version prompted participants to post hoc rationalize their ratings rather than providing genuine insight into the process they used for making choices. That would explain why the ostensibly objective (and hence less socially undesirable) options – the occupation's gender demographics, and physical requirements – were by some distance the most frequently endorsed criteria.

[N] [S] **Concluding thoughts** When we started this project, we were motivated by a worry about the use of demographic gender proportions as a proxy for gender stereotypes. Our worry might have been slightly overblown: Insofar as we could evaluate, the percentage of women working in an occupation was highly correlated with how stereotypically female it was rated by our participants. For occupations, for which we had percentages, such percentages apparently would make fine proxies.

That being said, actual ratings of gender stereotypes are still to be preferred. For one, imperfect correlations – in our case: correlations between .85 and .89 – leave room for error. Particularly for individual occupations, percentage scores can be misaligned with gender stereotypes (e.g., 85% of "tailors" in the UK demographic data were female, and yet the occupation was rated as stereotypically male by our UK respondents; $M_{6point} = 2.57, M_{7point} = 2.92$). For another, unambiguous demographic data was available only for about a third of our occupations. We had to make several subjective choices in compiling relevant gender percentages (e.g., about how to map our occupational labels to governmental job categories; or about whether to exclude occupation counts that the UK Office for National Statistics deemed potentially unreliable due to small sample sizes – which we chose against, i.a.), which could affect results. These hassles and potential confounders can be avoided by simply directly asking participants about their gender stereotypical views. Or by getting others, like us or Social scientist collaborators, to collect them for you. Research questions about the impact of AI on people are inherently multidisciplinary. They require multidisciplinary collaboration.

Acknowledgements

The authors wish to thank Leendert van Maanen for his feedback to an earlier version of this draft.

Chapter 5

There was no Gender Stereotype Priming in our Lexical Categorization Task

Dominik Bachmann[1,2], Katrin Schulz[1], Leendert van Maanen[2]

[1] Institute for Logic, Language and Computation, University of Amsterdam
[2] Department of Experimental Psychology, Utrecht University

Abstract

When asked about their views on controversial topics – like the (personality) traits they associate with men or women – people might not always provide honest and straight-forward responses (e.g., to avoid offending someone). Implicit measures were designed to circumvent such problems: If people's views are not directly asked about but inferred from performance on a task, people might not be aware of what is being tested – and hence unable to warp their responses. Here we set out to test this assumption for one implicit measure, *sequential stereotype priming*. We tested whether evidence for gender stereotypes could be obtained in a novel design that made it particularly unlikely that participants knew of the study's purpose (i.e., we included a coverstory, and paid participants to guess the purpose of our study, and yet only 11 of our 172 participants guessed that the study was about gender), and controlled for potential confounds that may have effected previous gender stereotype priming studies (i.e., response priming, and default response tendencies). Pictures of people in different occupations served as primes, and participants categorized agency- and communion-related words either as adjectives or nouns. Across different statistical models, no matter whether we analyzed reaction time or accuracy, we found no evidence for gender stereotype priming. Our results add to a growing body of null and inconsistent findings and suggest that gender stereotype priming effects are elusive once one controls for alternative explanations.

5.1 Introduction

"Exactly which stereotypes do you hold?" and "To what extent do you hold them?". If people gave straight and accurate answers to these two questions, social scientists could do a lot with them. After all, what generalized beliefs people hold about different categories of people touches on several important topics, like how people perceive their environments – where it is often efficient to rely on mostly accurate stereotypical heuristics like "seniors tend to react slower" – or how different categories of people, guided by their views of each other, interact.

However, if you were to pose these questions, you could not necessarily expect to receive such straight and accurate responses. Firstly, people might be weary of coming across rude or bigoted when providing honest answering, opting instead to provide responses they deem more socially acceptable. Secondly, people might be unaware of which stereotypes they hold. Different versions of these potential issues with investigating stereotypes through self-reports – that responders might deliberately change their answers, or be incapable of accessing the mental content we wish to measure (see e.g., Cameron et al., 2012; Gawronski & Hahn, 2018) – have motivated research into the development of *implicit measures* (Corneille & Gawronski, 2024; Gawronski & Hahn, 2018).

"Implicit" here means that the assessed mental content – in our case: stereotypical beliefs – affects measurements we take of that person in an automatic manner (Gawronski & Hahn, 2018). If we observed the impact of such automatic effects, it would allow us to draw inferences about a person's stereotypical beliefs, without the person having to introspect about these beliefs – circumventing the potential issue of them being incapable of such correct introspections – and without them having the opportunity to deliberately influence the measurements we make of them.

But for an implicit measure to fulfill this role, two strong assumptions have to be met: Firstly, what we observe must unambiguously stem from an automatic effect of the to-be-measured stereotypical beliefs. Secondly, since implicit measures are "notoriously sensitive to faking strategies" (Corneille & Gawronski, 2024), participants must remain unaware that we are attempting to measure stereotypical beliefs, as this will make it unlikely that participants deploy deliberate response strategies (like faking) which systematically impact our measurements.[1]

In this paper, we test whether we can detect effects of stereotypical beliefs with one type of implicit measure, *sequential stereotype priming*, if our design meets these two assumptions. In sequential priming tasks, people are shown two stimuli, the *prime* and the *target*, and have to perform a task with the target

[1]This assumption is self-evidently critical also for alternative conceptualizations of implicit measures (e.g., Petty et al., 2008) that label measures as implicit or explicit depending on whether or not respondents are aware of what is being measured (cf. our process-based conceptualization, which is based on Gawronski & Hahn, 2018).

stimulus. If the prime stimulus is semantically related as opposed to unrelated to the target stimulus, performance on the task improves, usually becoming faster and/or more accurate (Kidder et al., 2018; Wentura & Degner, 2010). For example, participants might respond faster if the prime "bird" as opposed to the prime "table" precedes the target "beak", as birds and beaks are semantically related. Sequential priming paradigms (and arguably most other implicit measures; Gawronski & Hahn, 2018) are build on the idea that activation of a concept in memory can spread to other related concepts and that we can, through such involuntary spread activations, explore the mapping of concepts in people's minds. For example, if exposure to the concept "bird" indeed facilitated your ability to act on the concept "beak" that would imply that in your memory the two concepts are closely associated.

Many different tasks involving the target can be employed to find evidence for such spread activations, ranging from word completion tasks (e.g., "Fill in the missing letter: b_ak"), through lexical decision (e.g., "Is beak a word?") and pronunciation tasks (e.g., "Say the following word: beak") to semantic judgments (e.g., "Is beak larger or smaller than a table?"). *Sequential stereotype priming* studies employ such priming paradigms to explore associations that people have made between social groups and traits. For example, if after the prime "woman" you act faster and/or more accurately on a task involving the concept "caring", this social group and trait are apparently strongly associated in your memory.

Assumption 1: Observed effects stem from stereotypical beliefs In its strongest form the first assumption cannot be met; we cannot know unambiguously that it is the mental content we wish to measure which led to observed behaviors rather than any other unobservable construct. Nevertheless, there are at least two important potential confounds to consider to make apparent evidence for stereotype priming less ambiguous. Firstly, so-called *response priming* is a concern: If the experimental task can be applied also to the prime stimulus, processing of the prime can prepare a response to the task, which has to be inhibited with incongruent but not with congruent target stimuli. Then, at least part of the observed apparent priming effect can stem from response competition (similar to a Stroop effect; Wentura & Degner, 2010) rather than from the semantic relationship between target and prime. For example, in a study that uses photos of people as primes and involves categorizing trait words (e.g., brave, modest, etc.) as either stereotypically feminine or masculine, apparent priming effects – that male faces facilitate the categorization of masculine trait words – could simply stem from male faces preparing the "masculine" response, a response that can be immediately executed for masculine trait words but must be inhibited for feminine ones.

This is a serious confound, as, compared to the effects of genuine semantic stereotype priming, the effects of response competition appear large. For exam-

ple, Tsamadi et al. (2020) could find evidence for priming effects when using a response priming paradigm, but no evidence for priming when using a design with identical trial structure and stimuli but a task that precludes response priming. Additionally, the meta analysis of Kidder et al. (2018) found that stereotype priming designs which allow for response priming have substantially larger effect sizes than those that preclude it.[2]

As causal attribution is difficult with paradigms that allow for response priming – it is unclear how much of an observed effect reflects stereotype priming as opposed to response competition – we believe that experimental designs should preclude response priming effects. As Kidder et al. (2018) found no evidence of priming for two of the most commonly used tasks that preclude response priming, the lexical decision task and the pronunciation task, they encourage the use of *semantic categorization tasks* (i.e., tasks that involve participants sorting target stimuli into categories, based on their meaning), for which they did find an effect.

Two such semantic categorization tasks provide examples for a second important confound of apparent priming effects: default responses. Both Wang et al. (2017) and Zhang et al. (2018) employed tasks that asked about stereotypical congruency. In Wang et al. (2017), participants observed the prime words "female", "male", "adjective" or "noun", and then saw a stereotypically female or male adjective or noun as a target word (e.g., "heroic" as a male adjective, or "nurse" as a female noun). Participants' task was to indicate whether or not the prime and target words were congruent. Depending on the prime, congruency was relative to gender stereotypes or to lexical categories – for example, "male" and "heroic" is congruent, while "adjective" and "nurse" is incongruent. In Zhang et al. (2018), participants saw pictures of female or male faces as primes, and then pictures of tools or of kitchenware as targets, and their task was to determine whether the prime and target were congruent in gender stereotypes (e.g., a female face followed by a ladle is congruent). Both studies found that participants responded more accurately and more quickly on congruent trials.

While neither of these studies employed designs prone to response priming – the question "are prime and target congruent?" cannot be applied to the prime in isolation – response competition is nevertheless a potential confound: The same response pattern could be found if "congruent" was participants' default response, which would lead to slower responses and more mistakes whenever participants have to deviate from that default. Priming designs have to hence preclude that default responses or similar response tendencies could lead to hypothesis-consistent effects. In our design, we did so by ensuring that both our response options were

[2]Kidder et al. (2018) concluded that there overall still was evidence for stereotype priming amongst studies with designs that preclude response priming. In our view, that claim was premature given the large uncertainty around their estimated effect size, $d = 0.1$ $[CI\text{: } 0.01, 0.2]$, and given that their conclusion hinges on their funnel plot-derived judgment that the estimate was not inflated by publication bias. (For a critique of funnel plots as a diagnostic tool for publication bias see e.g., Simonsohn, 2017).

equally often the correct and the incorrect response (see Section 5.2.2), though other approaches are possible.

Assumption 2: Participants are unaware of what we aim to measure In addition to the potential confound of a default response option, in Wang et al. (2017) and Zhang et al. (2018) participants had to explicitly assess gender stereotypes, making it quite obvious what researchers aimed to measure. While few tasks explicitly mention stereotypes, thinly concealed measurement aims are also concerns for other gender stereotype priming studies whose designs preclude response priming. For Müller and Rothermund (2014) all, and for Banaji and Hardin (1996) two thirds of prime words were related to gender. Additionally, no explanations or cover stories were provided for why these prime words were shown, potentially drawing more attention to their gender associations with the targets. Additionally, half of Müller and Rothermund (2014)'s participants performed a gender stereotype response priming task (i.e., had to categorize names as male or female after being primed with stereotypically male and female attributes) before the semantic categorization task. To those participants, the goal of measuring gender stereotypes likely was even more obvious.

Müller and Rothermund (2014) did not collect this data, but for Banaji and Hardin (1996) indeed 7 of their 60 participants reported awareness of a possible gender relationship between primes and targets. This might have been an underestimation, as participants might have chosen not to reveal that they saw through the manipulation to avoid "deliver[ing] the bad news to the experimenter" (Karylowski et al., 2001), or for other reasons like wishing to leave the experiment as quickly as possible instead of having to talk with the researcher.[3]

To more effectively conceal the purpose of the study than did Banaji and Hardin (1996), Karylowski et al. (2001) developed a clever design that provides participants with a rationale for why primes are presented. Participants were invited to a study that purportedly investigated short-term memory in situations where people need to do "two things at once". As primes, they used trait adjectives and job titles with male, female, or neutral gender stereotypical associations. Subsequently, they categorized target words either as names or non-words. After receiving feedback to their response, participants were then asked a question about the prime word: Participants in the semantic condition were asked whether the word was semantically related to a different word (e.g., "Was the initial word related to "happy?"), while participants in the structural task were asked whether

[3]While Karylowski et al. (2001) bring up that participants might politely withhold their awareness of a manipulation, they report that none of their participants revealed awareness of the study's purpose, themselves blindly trusting that there are no "polite withholders" in their study. One way to avoid potential pitfalls, like participants providing socially desirable responses to the question of what one tried to research and them wanting to move on from an experiment instead of proving such responses, is to provide financial incentives for participants to respond correctly. We employ this approach in the current study.

the prime word included a particular letter. This question about the prime provided participants with a reason for why a prime was presented, while allowing Karylowski et al. (2001) to manipulate participants' attentional focus from or to the meaning of the prime word.

Additional desiderata As Karylowski et al. (2001)'s design came closest to meeting our desiderata – for a stringent test of sequential stereotype priming that precludes alternative explanations like response priming while concealing the purpose of the study from participants – our design builds on theirs.

To make it more unambiguous that effects stem from stereotypical beliefs, we changed Karylowski et al. (2001)'s manipulation of attentional foci to be within-subjects. Compared to between-subjects differences, we believe that we would obtain conceptually stronger evidence for stereotype priming, if the same participants show evidence of strong priming while attending to gender stereotype-relevant information but less so while attending to other information[8] provided by the same prime (besides, this way we can ensure that participants attend to the relevant semantic dimension of the prime – here: gender – which some authors claim to be crucial; see e.g., Spruyt et al., 2009). Then differences in effects cannot stem from subtle differences between experimental setups or between participants.

We made three further changes that in our view make it more valid to infer that gender stereotype priming took place if an effect is observed: Firstly, we significantly decreased the time between prime presentation and task response, from Karylowski et al. (2001)'s 2000ms to 50ms. With shorter *stimulus onset asynchronies* alternative explanations, of effects stemming from controllable rather than automatic processes, become less likely (Wentura & Degner, 2010). Secondly, while words are currently the default prime choice for most experiments, we agree with Tsamadi et al. (2020) that pictures of people should be used as much as possible when investigating stereotypes. If we want to investigate which stereotypes are automatically activated when people perceive other people, we ought to give them people to perceive – not just verbal labels like "nurse". Consequently we used pictures of people as primes.

Thirdly, to tie in with existing social psychological literature on gender stereotypes, we use existing lists of terms from the "big two" of personality traits (Abele & Wojciszke, 2014), agency and communion. These two dimensions are highly associated with masculinity and femininity, respectively (see e.g, Abele & Wojciszke, 2007; Ellemers, 2018), and currently represent the main way of conceptualizing the content of gender stereotypes (Heilman et al., 2024). By using these words as targets, we ensure that the concepts that are activated by seeing pictures of people are actual meaningful stereotypes.

Statistically, we make a few improvements, as well, all with the goal of making our data more informative: Firstly, we decided to analyze both accuracy and

reaction time data. Despite the fact that priming-consistent improvements in task performance can be apparent both in faster as well as in more accurate responses (Cameron et al., 2012; Gawronski & Hahn, 2018), and despite the fact that accuracy-based priming effects have been argued to be more reliable than those based on reaction times (Cameron et al., 2012), most studies only try to detect priming effects with linear models on reaction times of correct responses (e.g., in Banaji & Hardin, 1996; Karylowski et al., 2001; Müller & Rothermund, 2014; Tsamadi et al., 2020; White et al., 2018, i.a., but not e.g. in Wang et al., 2017; Zhang et al., 2018). This throws away information stemming from participants' accuracy, as well as from reaction times on trials on which participants gave incorrect responses.

Finally, to avoid ceiling effects and hence make accuracy data more informative, we employ a categorization task that is harder than Karylowski et al. (2001)'s lexical decision-like task (the switch came also in light of the many null findings associated with lexical decision tasks; see e.g., Huang & Rothermund, 2022; Kidder et al., 2018; White et al., 2018) and remove their trial-by-trial feedback, which they included to make participants more accurate.

5.2 Method

5.2.1 Participant recruitment

Our here-described recruitment plan and exclusion criteria (as well as all analyses reported in Section 5.3.1) were preregistered in https://aspredicted.org/7pd7dz.pdf. We collected data on Pavlovia (https://pavlovia.org/), a website for hosting PsychoPy (Peirce, 2007) experiments, and recruited on the participant-recruitment website Prolific (https://www.prolific.com/). The experiment was open to participants that had not partaken in one of our pilots for this experiment, were located in the USA, fluent in English, aged 18-55, and not diagnosed with Dyslexia or colorblindness, and had an Prolific approval rate of above 97% (i.e., the percentage of times past researchers accepted a participant's work, as opposed to refusing to pay them; e.g., because they did not correctly perform a study). Participants were paid about 8 euros (the exact amount depending on how much of a bonus they won, when guessing the purpose of our study) for completing the (median) about 55 minutes-long experiment.

We collected data for 100 participants that did not meet any of our exclusion criteria. That final sample had a mean age of 37.82 years (SD = 9.28) and contained 46 women, 52 men, and 2 people that chose "other", when asked about their gender.

Criteria for excluding participants We preregistered that we would exclude participants if they either a) performed poorly on our main experimental task

(i.e., their accuracy on the *word categorization task* was below 60% on 5 or more of the 16 experimental blocks), b) had 20% or more of their trials excluded for responding too quickly or too slowly, c) in our question of what the experiment is about, came close to correctly guessing the hypothesis or d) noted in their response to that question that (some of) our pictures were AI-generated.

Forty-five participants were excluded for performing too poorly, 4 for having too many trials excluded, 12 for meeting both these criteria (i.e., poor performance and too many excluded trials), and 11 for correctly guessing the hypothesis of our study. In total, we hence collected data from 172 participants to reach our sample of 100 non-excluded participants. Within two weeks, we recruited all participants across four waves of testing. These were necessary, as we had to repeatedly check how many participants would need to be excluded and hence re-recruited.

Criteria for excluding trials Excluding trials on which participants responded too slowly and too quickly is recommended when modeling performance based on reaction times (Boag et al., 2025), as responses on such trials tend to come from different data-generating processes. For example, when responses are very fast, many stem from missclicks or random guesses rather than from participants' genuine attempts to complete a trial.

For slow responses, we excluded each trial whose reaction time was more than 3 inter-quartile ranges slower than their median response time – a conservative criterion which nevertheless led to the exclusion of about 3% of all trials in our final sample. For fast responses, we attempted to determine statistically at which response time participants started performing better than chance and excluded trials with response times faster than that value. Specifically, with ordinary least squares, we fitted a logistic curve that predicts the chance of a participant providing a correct response on a trial, based on the trial's reaction time. The logistic curve was fitted with a formula based on van Maanen et al. (2019)'s conditional accuracy function.[4]

In their application, parameter d represented the amount by which the logistic curve is shifted on the reaction time x-axis. We added a lower asymptote of .5 to their formula (i.e., as we had a binary task, we expect participants to have a

[4]

$$p(\text{correct}) = 0.5 + 0.5\left(\frac{1}{1 + e^{\,b(RT-d)+\frac{c}{RT-d}-a}}\right)$$

with $min(RT - d) \coloneqq 10^{-300}$, as the function is undefined for $RT - d = 0$, and illogical for $RT - d < 0$ (i.e., it would express that performance exceeds chance for a participants for reaction times longer than any that has been observed for them). a, b, c and d are fitted parameters; b is the rate at which accuracy decreases with larger reaction times (after initially increasing with them), larger a values increase the upper bound of the logistic curve (and hence the maximum accuracy a participant reaches) and larger c values decrease the steepness of the logistic curve, see van Maanen et al. (2019) for details.

performance floor of 50% correct responses). This changes the interpretation of parameter d to now represent the reaction time at which participants' performance starts exceeding chance (i.e., the reaction time at which the logistic curve starts rising above the asymptote, leading to a predicted accuracy above 50%).

Fitting curves and excluding trials like that works well for most participants, but not for those that have relatively constant levels of accuracy, as their best-fitting logistic curve is one that starts rising at very low reaction times – times faster than humanly possible genuine responses – to quickly reach their constant accuracy level. We hence decided to exclude participants' trials if their reaction time was lower than their d parameter value or lower than 300ms (a commonly used criterion for "too fast responses" on simple perceptual tasks; Boag et al., 2025, and hence a conservative criterion for our much more cognitively demanding categorization task) – whichever of the two values is higher.[5] Figure 5.1 depicts examples of these curves for three types of participants: a regular participant that had some excluded trials but was not excluded as a whole (Panel A), a participant that would have passed the accuracy inclusion criterion but was excluded for having too many too fast and too slow trials (Panel B), and a participant with a d estimate that was so low that the 300ms criterion was applied, instead.

As, to our knowledge, this way of excluding fast trials is novel, we preregistered to test whether a more traditional singular cutoff value of 500ms would lead to different conclusions than our per-participant-determined logistic cutoffs (see Section 5.3.1). Of participants that passed our accuracy inclusion criterion, applying the new criterion led to the exclusion – and hence recollection – of 4 participants (for one of them, see Panel B of Figure 5.1), while the 500ms cutoff would not have led to any additional exclusion. In our final sample of 100 participants, the new criterion led to the exclusion of 109 trials (i.e., 0.3% of trials) for being too fast.

5.2.2 Stimuli

Prime pictures The prime pictures were all possible combinations of one of 20 background shapes and one of 20 foreground photos depicting a person. The combination of 10 colors chosen to be visually distinct (i.e., blue, brown, green, gray, orange, pink, purple, red, yellow, and white) and 2 shapes – circles or squares of equal surface area – made up all 20 different background shapes. The combination of 10 occupations and two genders – male or female – made up all 20 different foreground photos. Two example stimuli are depicted in Figure 5.2.

[5] A more principled approach would be to fit the logistic curves in a Bayesian hierarchical manner, as pooling information across participants makes the parameter estimates for d more accurate. This might have also removed the need for the secondary 300ms cutoff value, as extremely low estimates for d would be shrunk towards the center of the overall distribution of d estimates. We leave this topic to future research, as our attempts to hierarchically model this logistic curve with brms (Bürkner, 2017) failed.

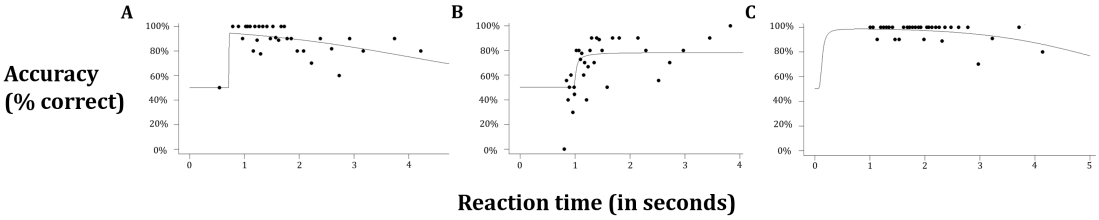


Figure 5.1: Three examples of the logistic curves we used to exclude trials that were too fast. Lines represent predicted accuracy based on reaction time, dots represent observed accuracies of trials binned by reaction time (e.g., the left-most dots represents a participant's accuracy across their 10 fastest trials, the second-left most dot their accuracy across their 11th to 20th fastest trials, etc.). The dots' location on the x-axis represents the bin's mean reaction time (e.g., despite panel C's leftmost dot being right of the 1 second x-axis tick, that bin contains individual trials with reaction times faster than 1 second). The reaction time at which a curve starts rising (after the flat line at accuracy = 50%) represents our cut-off: Trials faster than that value are excluded for being too fast (as the participant's accuracy does not exceed chance performance). **Panel A** depicts a participant with several excluded trials, but not enough for the participant to be excluded as a whole. **Panel B** depicts a participant who was excluded because too many of their trials were excluded for being too fast or too slow. **Panel C** depicts a participant whose performance was so consistent across reaction times that the curve starts rising very early. For such participants a 300ms cutoff was applied, instead.



Figure 5.2: Two examples of prime stimuli, depicting a male construction worker and a female soldier, respectively. The primes are combinations of a person in the foreground (one of 10 occupations, either as a woman or a man) and a shape in the background (one of 10 colors, either as a square or a circle).

The central "photos" were generated with Midjourney (https://www.midjourney.com), proprietary generative artificial intelligence software for image generation. Our instructions for creating the photos ("prompts") are listed in Appendix B.1.[6] Once we were satisfied with our revisions on Midjourney, we manually removed the backgrounds of the photos and cut them so that only torsos are visible. As is the case for all our other experimental materials, initial Midjourney-generated photos, edited versions without backgrounds, and all 400 final prime stimuli are accessible at https://osf.io/frq8c/overview?view_only=5ed5db2669b24b4e8e682b338b0ea382.

We chose to include 5 occupations that were stereotypically masculine (construction worker, firefighter, soldier, farmer and surgeon) and 5 that were stereotypically feminine (secretary, nurse, flight attendant, florist, and cleaner), as determined by a prior study (i.e., Chapter 4) in which we asked US citizens to rate occupations on whether they are seen as masculine or feminine in their culture.

Target words Target words were either agency-related, communion-related, or distractors. When choosing agency- and communion-related words, we prioritized words from Abele et al. (2008) and Abele et al. (2016) whose surveys identified words that validly represented positive agency and communion words (while controlling for psycholinguistic variables like word frequency) for the population we

[6]The AI-generated pictures seemed to represent people of those occupations well: Participants required an average of 43.28 trials (SD = 2.73) to complete the *occupation guessing task* (see Section 5.2.4) compared to the minimum number of 40 required trials, which means that they correctly picked the intended occupation out of four options in about 92% of trials. Participants also responded correctly on about 91% of questions about the occupations in the "person focus" blocks' *distraction task* (see Section 5.2.3), suggesting that they also perceived the photos as depicting the intended occupation during the subsequent main experiment.

tested: US Americans.

For agency- and communion-related adjectives, we hence exclusively used trait words from Abele et al. (2008). We chose against including negative trait words, like "boastful", which expresses a lack of communion (Abele et al., 2008), as we found it difficult to predict how priming would impact them (e.g., whether priming masculinity would activate such negative communion traits or merely activate positive communion traits less than does femininity). We further chose against differentiating subconstructs of agency and communion (e.g., the assertiveness vs competence dimensions of agency; Abele et al., 2016), as we did not find enough trait words to include an even number of words belonging to either subconstructs in adjectives, let alone in adjectives and nouns.

For nouns, we used words from three different sources: Firstly, we derived noun forms of trait words from Abele et al. (2008) and Abele et al. (2016) (e.g., turned "empathetic" to "empathy"), which were not already adjective target words. Secondly, we adapted nouns from Trapnell and Paulhus (2012)'s questionnaire on agency and communion-related values and life goals. Finally, we adapted nouns from Pietraszkiewicz et al. (2019)'s natural language processing-derived dictionary of agency- and communion-related words. All our agency- and communion-related target words, and where we sourced them in the literature are listed in Appendix B.2.

Our distractor words were meant to be unrelated to stereotype priming and were included to disguise the purpose of our study. We randomly sampled these words from the SCOPE psycholinguistic database (Gao et al., 2023) in a manner intended to control for psycholinguistic properties such as word length, familiarity and valence (for details, see Appendix B.3). We similarly created a word list for the practice trials (see Section 5.2.4) of our experimental task.

It is possible that such tight controls were not necessary. After all, the distractor words are not part of any statistical analysis and the psycholinguistic properties of the target words should be controlled for, as we obtain evidence for priming by contrasting accuracy and reaction times of the same words under different conditions (i.e., with congruent vs incongruent primes, and in a block that draws the attentional focus to or from the person on the prime picture; see Section 5.2.4). We however hoped that making the distractor words psycholinguistically similar to our stereotype ones would decrease the chance that participants create heuristics which impact their performance on hypothesis-relevant trials (e.g., "If the word is short, it's a noun") or that participants notice that some trials – the hypothesis-relevant ones – involve qualitatively different target words than others.

5.2.3 Procedure

Description of an experimental trial Each block started with a screen that told participants, through text and a small icon, whether they should attend to –

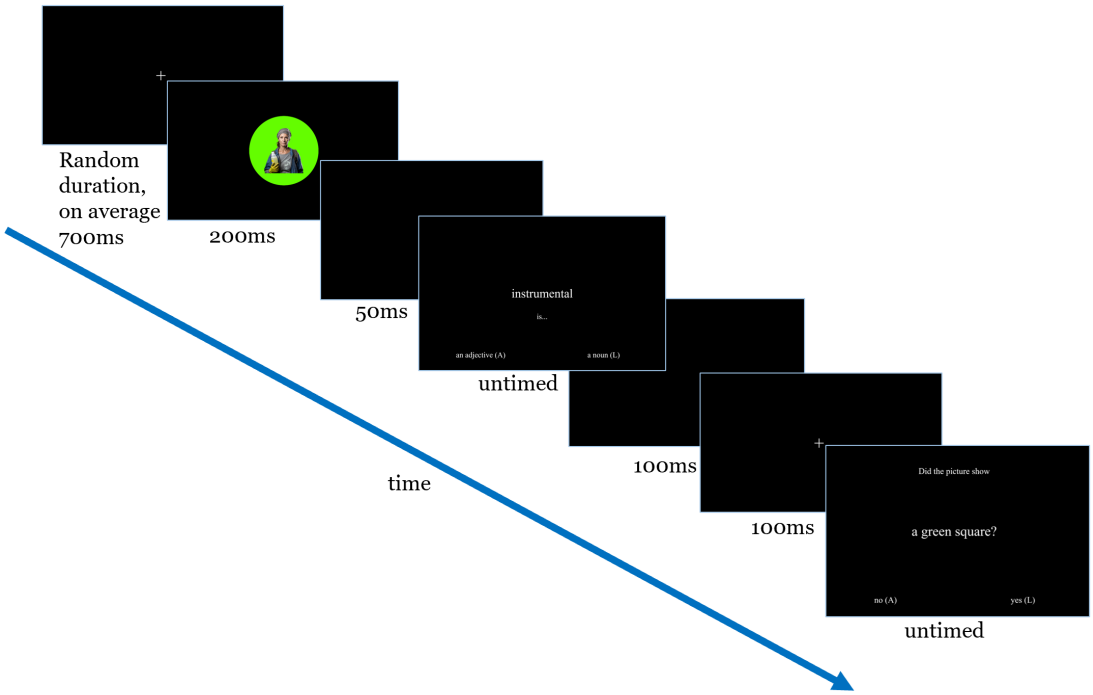


Figure 5.3: Timeline of an experimental trial from a "shape focus" block. The duration of the first fixation cross is drawn from an exponential distribution (for detail, see Footnote 7). The fourth screen depicts the *word categorization task* with a distractor target word (i.e., "instrumental" is neither agency-related nor communion-related). The last screen depicts the *distraction task* where the proposed description is incorrect on the 2-level factor (background shape).

and consequently whether they would be asked about – either the prime picture's person or its shape. Whenever they felt ready, participants could then begin a block of 20 trials by pressing a key. For reference, an example experimental trial from a "shape focus" block is depicted in Figure 5.3. Additionally, the PsychoPy script for the experimental task is accessible at https://osf.io/frq8c/overview?view_only=5ed5db2669b24b4e8e682b338b0ea382.

Each trial started with a black blank screen that was displayed for 200ms. Subsequently a white fixation cross was displayed on a black background for a random duration, the average duration being 700ms.[7] Then the prime picture

[7]The duration of the fixation cross was drawn from a truncated exponential distribution with a rate of 2, a lower bound of 0.2, and an upper bound of 5; a distribution suggested by Boag et al. (2025). This leads to an average duration of 700ms with a standard deviation of 500ms and a strong positive skew. Having a hard-to-predict trial onset is generally recommended, when one attempts to model data with evidence accumulation models (see e.g., Boag et al., 2025). As evidence accumulation starts at the presentation of the target word, this would have involved choosing a random duration for the blank screen between prime and target. We chose against that, as long durations between prime and target (which would be sampled on

(see Section 5.2.2) was presented for 200ms, followed by a short 50ms black blank screen that we included to decrease the impact that visual aftereffects of the prime picture have on the processing of the target word.

Subsequently, the target word was presented in white letters on a black background to the participant alongside the instruction to classify the target word as an adjective or a noun. This *categorization task* had no response deadline; participants could respond as slowly as they wanted to – though we did initially instruct them to respond as fast and as accurately as possible. After providing a response, the presentation of the target word concluded. We did not record responses nor proceeded in the experiment if a response occurred within the first 50ms of target presentation, as only accidental key presses could lead to such fast responses. A quarter of target words were related to agency, a quarter to communion and half were distractors intended to hide the purpose of the study. This meant that a quarter of trials included congruent prime-target pairings (i.e., female picture and communion word, or male picture and agency word), a quarter incongruent ones (e.g., male picture and communion word) and half included stereotype-unrelated pairings.

Per participant, each occupation was randomly assigned one noun and one adjective each for agency and for communion. Each participant hence responded to all our agency- and communion-related words, but participants had different occupation-word pairings. As there are two different photos per occupation (i.e., one per gender) and two different types of block (i.e., "person focus" vs "shape focus"), each target word was presented four times. This way we could contrast responses on the same word across prime genders and block types, while keeping equal the number of trials on which "adjective" and "noun" are the correct responses, respectively.

After providing a response to the categorization task, a blank screen and a fixation cross were displayed, each for 100ms. Subsequently, participants were asked a question about the prime picture with binary "yes" and "no" response options. Like in the categorization task, this *distraction task* (see next Section) was untimed and responses were not recorded – and hence a new experimental trial was not initiated – if participants pressed a response key within 50ms of the presentation of the question.

For both the categorization task and the distraction task, "a" and "l" were used as response keys. Across participants we randomized which of the two keys represented "adjective" vs "noun" and "yes" vs "no", respectively: Twenty-seven participants had the "a" key bound to "adjective" and "yes", 21 to "adjective" and "no", 30 to "noun" and "yes", and 22 to "noun" and "no".

---

some trials) could lead to alternative explanations for congruency effects that are unrelated to priming (Wentura & Degner, 2010). The best solution we could think of was to present the prime picture only for a short duration, and present the target very shortly afterwards. This way, evidence accumulation still starts very shortly after the random trial onset (i.e., shortly after the cross fades).

Distraction task In the distraction task, participants were asked a binary question about the prime picture in the form of "Did the picture show a [...]?". Depending on the type of block, these questions addressed either the person or the shape in the prime picture. In half of trials, an accurate description was presented. For example, if the prime picture depicted a female firefighter on a yellow circle, the question would be "Did the picture show a female firefighter?" in a "person focus" block or "Did the picture show a yellow circle?" in a "shape focus" block.

The other half of trials involve inaccurate descriptions of the prime picture. Specifically, in a quarter of trials, the description of the prime was wrong on the two-level factor that is relevant to this block: In the "person focus" block it would ask about the correct occupation but wrong gender; in the "shape focus" block about the correct color but wrong shape. In the remaining quarter of trials, the description of the prime was wrong on the ten-level factor that is relevant to this block: In the "person focus" block it would ask about the correct gender but about one of the other 9 occupations; in the "shape focus" block about the correct shape but about one of the other 9 colors. With this half/quarter/quarter split we ensured firstly that both response options are equally often correct and secondly that both dimensions of the prime picture (e.g., gender and occupation) are equally relevant to attend to.

Including the distraction task has three benefits: Firstly, it helps with disguising the purpose of this study. As the distraction questions ask about the picture, they provide a reason for why the pictures are included in the study (cf. studies were participants are asked to ignore the primes; see e.g., Banaji & Hardin, 1996). The questions are also consistent with the cover story that we used during recruitment that this experiment investigates "multi-tasking" between a linguistic and a visual memory task, two tasks that ostensibly use different brain areas. Secondly, the tasks helps us shape the processing of participants. Instead of participants freely associating about the photo – for example that the depicted person reminds them of a friend – they are nudged to semantically process the photo in terms of gender and occupation, as they will be asked questions about these two dimensions. Thirdly, the distraction task provides us with an opportunity to stringently test whether semantic priming took place: The distraction question manipulates how much participants semantically process the person on the prime picture. We should hence see a stronger priming effect in blocks in which participants primarily semantically process the prime's person rather than its shape.[8]

[8]Initially we designed the experiment to test for a full double dissociation of "no priming in shape focus blocks" and "priming in person focus blocks" which would present the most stringent test of (and hence potentially the most convincing evidence for) stereotype priming. However, upon piloting the study ourselves we noticed that it is quite difficult to exclusively attend to the shapes in shape-focused blocks, as the centrally depicted people occasionally capture one's attention. Follow-up studies might try to blend photos more with the backgrounds to make

5.2.4 Overview of the study

After providing informed consent, and entering their age and gender, participants performed an *occupation guessing* task with the photos of people with different occupations that are part of the prime pictures (see Section 5.2.2). Our goal for that task was to strengthen participants' association between an occupation and the photos that we intended to represent the occupation.

Specifically, a photo was shown for 400ms and subsequently participants had to choose, without speed pressure, which of four occupations – the correct one and three randomly chosen ones from our list of included occupations – the person on the picture had. On each trial, each picture can be depicted, unless it has been correctly identified twice. Consequently the task concludes when every picture has been correctly labeled twice. Participants received feedback on each trial: A 300ms long green "CORRECT" if they answered correctly, or a 1500ms long red "WRONG!" alongside corrective feedback (e.g., "The correct response was nurse!") if they answered incorrectly. As corrective feedback took longer, participants were incentivized to perform to the best of their ability.

After the occupation guessing task, participants were instructed on how to do the experimental task and subsequently performed 4 mini blocks of 5 practice trials, each. These practice trials were identical in setup to the main experimental trials (see Section 5.2.3) except that participants now received feedback for each of their responses to the categorization task and the distraction task: either a 300ms green "CORRECT" or 1500ms of a red "WRONG!" alongside corrective feedback like "crunchy is an adjective!" or "The picture showed a red square!".

Half of the mini practice blocks had the "person focus" and half the "shape focus" and the order of block foci was randomized. Like in the main experiment, there were waiting screens between experimental blocks which informed participants about whether the next block involved questions about the prime's person or shape, and allowed them to choose when to proceed to the next block of trials. For the practice trials, we used a list of words that was randomly sampled to be psycholinguistically similar to the target words (for more details, see Appendix B.3) and that did not overlap with the distractor words or agency or communion words that we used in the main experiment.

After completing the practice trials, participants could proceed to the main experiment, starting their first of 16 blocks of 20 trials. We pseudo-randomized the order of trials within block type to ensure that no subsequent trials depict the same target word. Subsequently we split the 160 trials of each block type into 8 blocks, each, and randomized the order of the resulting 16 blocks.

After completing all experimental trials, participants were told that they had the chance to win a bonus of up to 1 euro, depending on how close to correct their

the "tuning out" of people simpler in the "shape focus" blocks. We decided against that, here, as it would make the pictures less naturalistic and, in our estimation, hence less likely to automatically activate stereotypical associations.

guess was about what we tried to research. In a subsequent window they then got the chance to venture a guess, the exact instruction being "Please enter your best guess(es) for what this experiment is about (e.g., the exact hypotheses we try to answer).". All but one participant ventured a guess. We read all guesses and graded them by how close to the truth they were, rewarding them in increments of 25 cents. Participants that guessed correctly enough to receive 75 cents or 1 euro were excluded (see Section 5.2.1). Subsequently, participants were debriefed and the experiment concluded.

5.3 Analysis

Here we describe all preregistered analyses that we performed. For an overview of preregistered analyses that we forewent and why we did so, see Appendix B.4. All our data and analysis scripts are accessible at https://osf.io/frq8c/overview?view_only=5ed5db2669b24b4e8e682b338b0ea382.[9] As preregistered, we ran all here-reported analyses twice, once for trials and participants excluded based on the logistic curve reaction time criterion, and once with exclusions based on a 500ms reaction time criterion (see Section 5.2.1 for a discussion of these criteria). Here, we will only report the results based on the logistic exclusions, as across the two criteria the same models won the model comparisons (i.e., conclusions were identical).

In Appendix B.5, we provide results for a logistic regression and a linear regression to evaluate whether there is evidence for priming in our participants' accuracies and/or reaction times. Some science communicators argue that researchers should also report results for such "simplest valid analyses" (e.g., notably Greenberg, 2024, who coined that phrase) that one could use to test one's hypotheses, for example to make analysis sections easier to follow for readers. We will note however that for our data the here-described (generalized) linear mixed models are simply superior versions of the there-reported linear and logistic regressions.

5.3.1 Preregistered analyses

To evaluate whether stereotype priming took place, we fitted frequentist linear mixed models to participants' log-transformed reaction times[10] and generalized

[9] This includes data from two pilot studies. In the first pilot (called "E1" in the data), we tested 10 participants to see whether the experiment runs well. We noticed that performance was poor and consequently revised our instructions (e.g., to explain with examples what a noun and an adjective is). The second pilot ("E2") is fully identical to our main experiment ("E3"). We used the data from these 15 pilot participants to see whether performance improved to an acceptable level with the new instructions – it did – and to revise our exclusion criteria and analyses to finalize the preregistration.

[10] This transformation makes the distribution of reaction times less skewed, better satisfying the assumption of our linear models that residuals are normally distributed.

linear mixed models to their accuracy data (i.e., whether or not they categorized a word correctly) on relevant trials from the categorization task; that is: on all trials that a) were not excluded due to their reaction times (see Section 5.2.1) and b) involved target words that were not distractors (see Section 5.2.2). To find the model that best explains the data, we fitted different models that included the main effects of Block Type (i.e., whether people were asked about the person or the shape of the prime, this block), Congruency (i.e., trials are congruent if a female prime picture was paired with a communion target word or a male prime picture with an agency word, and otherwise incongruent), Lexical category (whether the target word is an adjective or noun) and Replicate (i.e., whether the current trial is the target word's first, second, third, or fourth presentation) as well as all their interaction terms. Models were compared with the Bayesian Information Criterion (BIC).

Model selection procedure We first simplified the random effects structure of our two preregistered full factorial hierarchical models until they converged. To do so, we followed Bates et al. (2015)'s procedure of iteratively using *Principle Component Analysis* (PCA) to identify random effects terms that can be excluded without notably decreasing model fit.[11] This way we identified a model each for reaction time and for accuracy (see the 0 columns of Tables 5.4 and 5.5) which could be fit without convergence issues or degenerate parameter estimates.

Subsequently, we simplified these models until no further exclusion of a predictor was supported by the data. Specifically, we iteratively identified for the current best model which predictors could be validly removed, fitted all of these possible simpler models and then identified a new best model, based on BIC, which we would attempt to simplify in the subsequent round of model comparisons. Fixed effects were valid targets for removal, if no higher-order interaction effect contained it (e.g., if a model included the predictors BlockType:Replicate and BlockType:Congruency:Replicate, only the latter could be deleted, because the 3-way-interaction term includes the 2-way-interaction term). For random effects, correlation terms (e.g., between a random intercept and a random slope) were removed before their random intercepts or slopes. Finally, if a fixed effect was removed, all their random slopes were removed as well (e.g., if the main effect for Congruency is removed, so are all random slopes for Congruency).

When no further simplification of models was supported, we attempted to increase the models' complexity until no further inclusion of predictors was supported by the data. Specifically, we iteratively identified for the current best

[11] For example, in the second iteration of simplifications for our accuracy model, we had 6 random effects terms for the grouping factor "participant" (1 random intercept, 2 random slopes and the 3 correlations between those slopes and the intercept). As the PCA indicated that one principle component explains all of the grouping factor's variance and that the random intercept near-perfectly loaded on that component, we removed all participant-related random effects except for the random intercept.

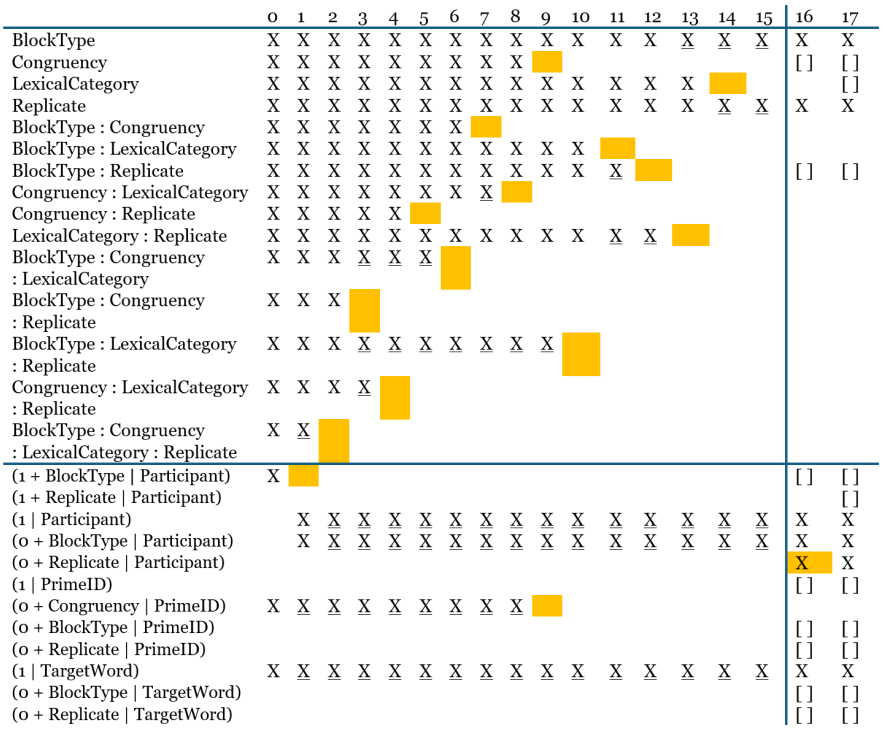
model which singular predictors could be added, fitted each model with a singular such inclusion and then identified a new best model based on BIC, for the next round of model comparisons. For fixed effects, we attempted to include each of our four main effect predictors that was currently not included in the model as well as each interaction term that was currently not included but whose component predictors were part of the current model (e.g., for a model with Congruency and Replicate fixed effects, the Congruency:Replicate interaction, the BlockType main effect and the LexicalCategory main effects would be tested). For random effects, all random slopes of included fixed effects and all not currently included random intercepts were tested for our three grouping factors "participant", "target word" and "prime picture". Additionally, if multiple random terms were included for the same grouping factor (e.g., a random intercept and a random slope), we also tested whether including their correlation would improve the model.

5.4 Results

We obtained no evidence for priming – neither for participants' reaction time, nor for their accuracy, as the winning models included no predictors that were related to the congruency between prime picture and target word.

For accuracy, the only supported fixed effect was lexical category: Participants performed better on adjectives than nouns (raw means: 88.6 vs 64.7%). For log-transformed reaction time, the winning model included main effects for block type – participants were faster in blocks in which they focused on the person rather than the shape (raw means: 2.19 seconds vs 2.29) – and for replicate; participants responded faster the more frequently they had seen a word (raw means, in seconds: $M_1 = 2.60, M_2 = 2.27, M_3 = 2.11, M_4 = 1.99$).

Besides none of the winning models including predictors related to Congruency, the sequence of model changes also suggests that no priming took place: Except for two times (i.e., random effects unrelated to priming were excluded in Generation 1 of the RT models, see Figure 5.4, and in Generation 5 of the accuracy models, see Figure 5.5), model comparison always favored models that had fewer predictors related to Congruency/priming (i.e., in both cases the inclusion of Congruency-related predictors was never supported and the main effects of Congruency were removed before any effects – including even the 3-way-interactions – of the remaining main effects; see the Generation 9 columns of Figures 5.4 and 5.5). Since all included predictors had one degree of freedom, BIC only differentiated based on model fit; so evidently priming-related predictors explained less variance than any other predictor.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
BlockType	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Congruency	X	X	X	X	X	X	X	X	X								[ ]	[ ]
LexicalCategory	X	X	X	X	X	X	X	X	X	X	X	X	X	X				[ ]
Replicate	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
BlockType : Congruency	X	X	X	X	X	X	X											
BlockType : LexicalCategory	X	X	X	X	X	X	X	X	X	X	X							
BlockType : Replicate	X	X	X	X	X	X	X	X	X	X	X	X					[ ]	[ ]
Congruency : LexicalCategory	X	X	X	X	X	X	X	X										
Congruency : Replicate	X	X	X	X	X													
LexicalCategory : Replicate	X	X	X	X	X	X	X	X	X	X	X	X	X					
BlockType : Congruency : LexicalCategory	X	X	X	X	X	X												
BlockType : Congruency : Replicate	X	X	X															
BlockType : LexicalCategory : Replicate	X	X	X	X	X	X	X	X	X	X								
Congruency : LexicalCategory : Replicate	X	X	X	X														
BlockType : Congruency : LexicalCategory : Replicate	X	X																
(1 + BlockType | Participant)	X																[ ]	[ ]
(1 + Replicate | Participant)																		[ ]
(1 | Participant)		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
(0 + BlockType | Participant)		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
(0 + Replicate | Participant)																	X	X
(1 | PrimeID)																	[ ]	[ ]
(0 + Congruency | PrimeID)	X	X	X	X	X	X	X	X	X									
(0 + BlockType | PrimeID)																	[ ]	[ ]
(0 + Replicate | PrimeID)																	[ ]	[ ]
(1 | TargetWord)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
(0 + BlockType | TargetWord)																	[ ]	[ ]
(0 + Replicate | TargetWord)																	[ ]	[ ]

Figure 5.4: This table indicates which predictors were included in the winning model of each **model comparison** generation for **log(ReactionTime)**. Each row represents a predictor. Each column represents a test generation. Generation 0 represents the most complex model that converged (see Section 5.3.1). In Generations 1 to 15, the exclusions of predictors were tested. In Generations 16 and 17, the inclusions of predictors were tested. Cells marked with X index predictors that were included but not tested, in this generation of models. Cells marked with X and [ ] index predictors whose exclusions and inclusion, respectively, were tested but not supported, this generation. Orange cells indicate changes which were tested and supported, this generation (i.e., the exclusion or inclusion of this predictor resulted in the model with the lowest BIC of this generation's models). The overall winning model is the model of Generation 16.

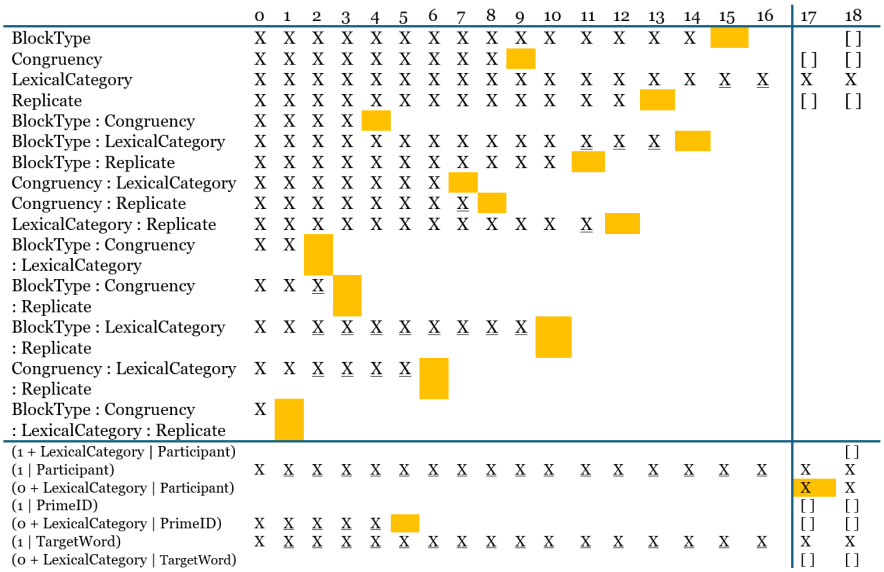


Figure 5.5: This table indicates which predictors were included in the winning model of each **model comparison** generation for **Accuracy**. Each row represents a predictor. Each column represents a test generation. Generation 0 represents the most complex model that converged (see Section 5.3.1). In Generations 1 to 16, the exclusions of predictors were tested. In Generations 17 and 18, the inclusions of predictors were tested. Cells marked with X index predictors that were included but not tested, in this generation of models. Cells marked with X and [ ] index predictors whose exclusions and inclusion, respectively, were tested but not supported, this generation. Orange cells indicate changes which were tested and supported, this generation (i.e., the exclusion or inclusion of this predictor resulted in the model with the lowest BIC of this generation's models). The overall winning model is the model of Generation 17.

5.4.1 Exploratory analyses

We conducted several exploratory analyses to explore alternative explanations of our results and caveats to our conclusions. Since in our main analyses we found no differences in conclusions based on which reaction time-based exclusion criterion (see Section 5.2.1) was applied, we performed these exploratory analyses only on the data where exclusions were based on the logistic criterion.

BIC punishes models quite severely for complexity. Would AIC lead to different conclusions? We saw two different ways of addressing this question: firstly, to compare all already fitted models with AIC and secondly, to rerun the model selection procedure outlined in Section 5.3.1 but this time guided by AIC rather than BIC.

When comparing the 107 already-fitted reaction times models (see Figure 5.4) the same model that had won based on BIC also won based on AIC. For the 85 accuracy models (see Figure 5.5) a model from generation 17 won the model comparison, meaning that model comparison based on AIC supported the inclusion of more predictors (unlike BIC which supported no changes after generation 16). We hence continued testing the inclusion of predictors in the way outlined in Section 5.3.1 but now based on AIC. This led to three more generations of adding predictors, but those were all nuisance factors related to Replicate.[12] In other words: Once more we found no evidence that stereotype priming took place.

When re-running the entire model selection pipeline with AIC the resulting models included many more predictors than the ones found based on BIC. The conclusion remained the same, however: there was no evidence for congruency influencing reaction time or accuracy. For reaction time, the winning model did not include any predictors involving Congruency. For accuracy, the winning model did not include the hypothesis-relevant BlockType:Congruency interaction effect, and the estimate for the Congruency main effect was virtually zero, $\beta = 0.0002, p = .997$.[12]

Your prime photos (are AI-generated), maybe some of them lead to abnormal priming effects or did not capture the profession well? We

[12]The predictors of the winning models were:

Accuracy, comparison by AIC: LexicalCategory + Replicate + (1 + LexicalCategory + Replicate | Participant) + (1 | TargetWord)

RT, comparison by AIC: see column 16 of Figure 5.4,

Accuracy, model search by AIC: BlockType*LexicalCategory*Replicate + Congruency + Congruency:LexicalCategory + Congruency:Replicate + Congruency:LexicalCategory:Replicate + (1 + LexicalCategory * Replicate | Participant) + (1 | TargetWord),

RT, model search by AIC: BlockType * LexicalCategory * Replicate + (1 + BlockType + LexicalCategory + Replicate || Participant) + (1 | TargetWord).

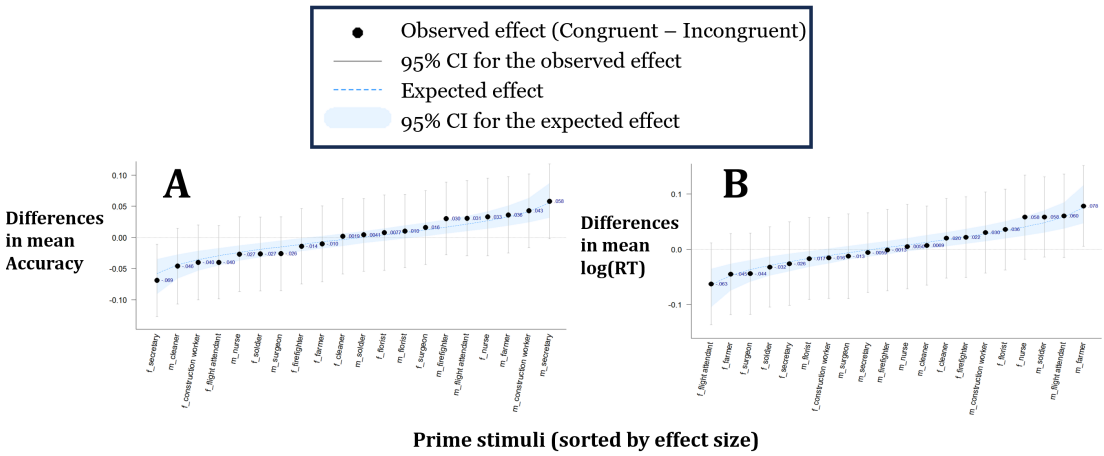


Figure 5.6: These stimulus plots depict on the y-axis mean differences between congruent and incongruent trials on accuracy (**Panel A**) and reaction time (**Panel B**), respectively, across the different prime photos. Prime photos are sorted on the x-axis by observed effect size. Black dots represent these observed effect sizes and gray whiskers represent their corresponding 95% confidence intervals (as determined by t-tests per prime photo). The blue dashed line indicates which effect sizes would be expected under the null hypothesis of all stimuli coming from the same distribution (in these cases: with mean effect sizes of ≈ 0). The blue band is the 95% confidence band of expected effects, as determined by 5000 permutation samples (for details on the method, see Simonsohn et al., 2025).

addressed this question with Stimulus Plots (Simonsohn et al., 2025), tools for testing whether all stimuli in one's study point towards the same conclusion or whether there are unexpected levels of hetereogeneity between stimuli. Figure 5.6 depicts such stimulus plots, which contrast the effects we would expect to observe with the effects we did observe across the photos which were part of our prime pictures (i.e., we pooled effects of prime photos across prime background shapes and colors). All observed effect sizes are within confidence bounds of what one would expect to see. In other words: There is no reason to suspect that any prime photo had an unexpectedly large or unexpectedly directed priming effect (e.g., that an occupation primed agency more for women than for men). Simonsohn et al. (2025)'s test of heteroegeneity also confirms that the observed differences in effect sizes are quite plausible under the null hypothesis of all primes being homogeneous, $p_{RT} = .90$, $p_{Accuracy} = .58$.

All observed effects are consistent with assuming that no priming took place. It is, for example, not the case that some stimuli, with strong hypothesis-incongruent priming effects, mask the hypothesis-congruent priming effects of other stimuli.

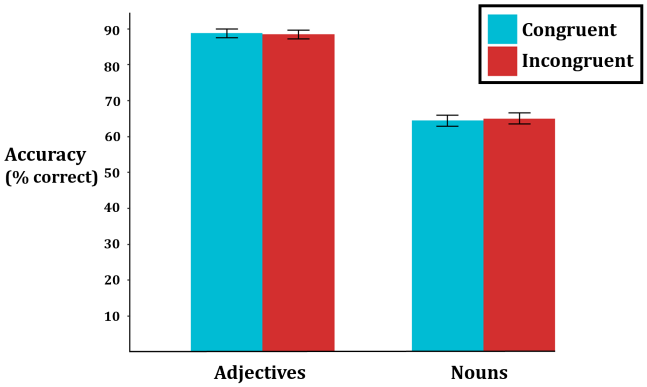


Figure 5.7: Mean accuracies of participants on congruent and incongruent adjectives (left plot) and nouns (right plot). Whiskers represent the Cousineau-Morey (Morey, 2008) 95% confidence intervals around the mean accuracy. (We apply a Morey correction for 4 factor levels, as there are 4 combinations of congruency and lexical category.)

Most priming studies use adjectives, maybe your nouns are the problem? We have no indication that adjectives and nouns differed in the extent to which priming took place. If such a trend existed and was of sufficient effect size, one of the winning linear models would have an interaction effect involving LexicalCategory and Congruency. Instead, as was the case for every other predictor involving congruency, such interaction terms were removed early in the model comparisons.

Also when looking at the raw data, we see no reason to suspect that any priming has taken place, selectively for nouns or adjectives. Across congruency conditions, participant have highly similar mean accuracies, within adjectives and within nouns (see Figure 5.7). And across congruency conditions, reaction time distributions are near-identical within (and, while that is not directly depicted here, also largely between) lexical categories (see Figure 5.8). It is, for example, not the case that for adjectives but not for nouns performance is faster on congruent trials.

5.5 Discussion

The goal of our study was to test whether evidence for gender stereotype priming can be obtained with a sequential priming design that precludes alternative explanations – like response priming and participants having default responses – and that successfully hides the purpose of the the study from the participants. In our study, no such priming appears to have taken place. This conclusion was robust across dependent variables – whether we tried to detect priming effects

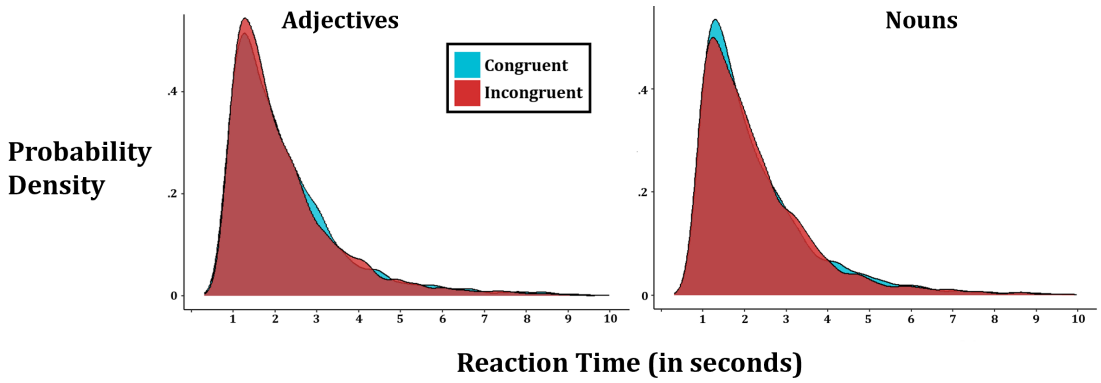


Figure 5.8: Kernel density plots of reaction time distributions for responses to adjectives (left plot) and nouns (right plot). We overlay reaction times for congruent trials (blue) with reaction times for incongruent trials (red), revealing that those are virtually identical. For this figure, we censored reaction times slower than 10 seconds. The density starts after 50ms, as faster reaction times were not registered (see Section 5.2.3).

on accuracy, reaction time, or both. It was also robust no matter whether one assumed that the gender of a prime picture or its gender-stereotyped occupation was more likely to activate stereotypical associations. Even when comparing only those trials that should have shown the strongest priming effects – when the prime's gender and occupation are both consistent with the stereotype word – with those that should have shown the weakest – when the prime's gender and occupation are both inconsistent with the stereotype word – we detected no evidence for priming.

That no stereotype priming took place in our study is in line with previous internally inconsistent (White et al., 2018) and null findings (Huang & Rothermund, 2022; Tsamadi et al., 2020) from studies that used the Lexical Decision Task. Contrary to Kidder et al. (2018)'s cautious optimism about semantic categorization tasks, we find our results also not terribly surprising in light of the results from previous studies using such tasks: Müller and Rothermund (2014) did not find evidence for gender stereotype priming in their high-powered, preregistered semantic categorization task. Karylowski et al. (2001)'s evidence for stereotype priming consisted entirely of "marginally significant" results, in a study of only 31 participants, for a (granted, theoretically motivated) subset of targets only. Additionally, since their design involved a very long stimulus onset asynchrony, effects other than automatically activated stereotypes could have brought about hypothesis-consistent results on the task (Wentura & Degner, 2010).

Banaji and Hardin (1996)'s "evidence for stereotype priming" amounted to the finding that participants are faster to characterize "he" and "she" as pronouns, following prime words that are gendered by definition. In our view, finding that words like "mother" and "waitress" have associations with the female pronoun

"she" is not evidence for stereotypical associations but of grammatical ones.[13] As we discussed in more details in the introduction, Wang et al. (2017)'s and Zhang et al. (2018)'s results could stem from participants having "congruent" as a default response, and both studies deploy paradigms that are transparently about gender stereotypes. Kawakami and Dovidio (2001) found evidence for racial stereotype priming, but in a design that thinly veiled the purpose of the study. But even if we take their results as convincing evidence for racial stereotype priming, gender stereotype priming might be harder to detect (consistent with e.g. Kidder et al., 2018, who found larger effect sizes for racial as opposed to gender stereotype priming).

It appears to us that gender stereotype priming effects are illusive, once one accounts for alternative explanations for them – at least when one tries to test for simple category-based (e.g., "female nurse") activations of stereotypes. The same conclusion might not hold for paradigms (e.g., Casper et al., 2010; Huang & Rothermund, 2023) which additionally aim to emulate the social contexts in which stereotyping occurs in real life. For example, Casper et al. (2010) found priming effects only when the social context was fitting (i.e., the stereotypical association between "Bavarian" and "beer" could be found in the context of a marquee but not in the context of a shop).

Limitations of our study and trade-offs we made While we were aware of such studies and find them promising due to their increased ecological validity, we chose against implementing such a design in our study. We feared that these tasks are somewhat unsubtle about what one tries to research: For example, after seeing the "Bavarian" prime word with a picture of a marquee and after subsequently reading the target word "beer", it should be easy to guess that the study is about stereotypes – at least for those participants (for whom we could find priming effects as they are) aware of the stereotype about Bavarians.

For a similar reason of believing their downsides to be too costly, our design also made some deviations from established best practices for sequential priming studies (as outlined e.g. in Spruyt et al., 2011; Wentura & Degner, 2010): Firstly, like all other stereotype priming studies we are aware of, we did not meet the "gold standard of semantic priming" (Wentura & Degner, 2010) of using every target word only once per participant, and hence did not eliminate the effects of target repetition. Given that our design required that the same occupation is paired with a target word category (e.g., agency noun) four times – across 2 attentional foci, times 2 primes of differing congruency – we did not have enough gender stereotype-relevant target words to use one per instance. Additionally, we did not consider it feasible to create a good Latin square design for words

[13] Banaji and Hardin (1996) also tried other prime words, including labels of occupations that were associated with a particular gender. For these more stereotype-related words, they found no effects.

across these four presentation instances and across occupations (especially in light of our intended sampling and exclusion plan). Instead we opted to randomly assign target words to occupations and used the minimum number, for our design, of 4 target repetitions. Additionally, we included "target word replicate" as a predictor in our models, statistically accounting for it. As we found it to be our single best predictor of reaction times, we believe that future studies should control for this, as well.

Secondly, we did not include a neutral priming condition, where the same target words are depicted without a prime that contains stereotype-relevant information. Including such a condition allows one to compute absolute sizes of priming effects by comparing primed versus unprimed trials with the same target stimuli (Wentura & Degner, 2010). It also allows one to test for the presence of situations where prime stimuli activate concepts counter to a stereotype (Huang & Rothermund, 2022): For example, encountering statements like "men aren't considerate" could paradoxically lead to an association between the concepts "man" and "considerate". When such associations exist and are activated by priming to similar extents as stereotypical associations (e.g., between "woman" and "considerate"), a lack of a difference in task performance between trials in which masculinity versus femininity are primed does not necessarily reflect the absence of gender stereotypes.

We chose against implementing neutral prime conditions (e.g., where background shapes without photos are shown as primes), as including neutral primes would have led to many more trials – when our experiment was already long – and to additional replicates of the target words, when we wanted to minimize their number. We deemed this a poor tradeoff, especially since Huang and Rothermund (2022) did not find evidence for such stereotypical and counter-stereotypical associations coexisting when they used neutral primes in their study of age stereotypes. We do not believe that this hampers the interpretation of our results much. We were interested in investigating the properties of stereotype priming as a measurement tool for gender stereotypes. Consequently, a hypothetical alternative explanation for our results of "femininity and masculinity activate stereotypical and counter-stereotypical concepts to similar extents" would be equally bad news for priming as a measurement tool as a conclusion of us not finding evidence for any activation. In either case we cannot use priming paradigms to measure people's gender stereotypes.

Two final potential caveats to our findings concern the nature of our task: Firstly, one could worry about the quality of our data, because online participants might have taken the task less seriously or been more distracted than participants in controlled laboratory environments. We do not anticipate that this substantially influenced our conclusion. While online experiments can lead to slightly smaller (see e.g., Chuey et al., 2024) and hence harder-to-detect effects, decreases in data quality relative to lab participants appear small (e.g., reaction time distributions can be recovered online; Sauter et al., 2022; though cf. Bürki, 2025 for

a more pessimistic perspective about online testing), especially for participants from Prolific (Uittenhove et al., 2023). Additionally, we implemented reaction time- and accuracy-based exclusion criteria to increase the quality of our data. This should filter out most careless online participants. Finally, as evidenced by their $\approx 91\%$ accuracy on distraction questions concerning the prime person, our participants perceived the stereotype-relevant content of a prime sufficiently for priming effects to take place.

Secondly, our task of categorizing words as nouns or adjectives could be a concern: Maybe priming more generally, not stereotype priming in particular, cannot be found with such a task. We agree that finding an absence of stereotype priming would have been more informative in the presence of regular semantic priming with the same paradigm (e.g., that participants perform better in categorizing "hospital" after seeing one of our "nurse" primes). We look to collect such evidence, next. If no kind of semantic priming can be found with our task, it could be due to participants not having to consider the meaning of words much beyond whether they are nouns or adjectives. We will note however, that our task arguably requires more consideration of the meaning of words than at least the commonly-employed "is it a word?" lexical decision task and Karylowski et al. (2001)'s "is it a name?" task. Both these tasks can be accurately performed solely by relying on recognition memory (i.e., basically by considering the question of "does this letter combination look familiar?") without considering the meaning of the targets.

A third concern could be our particular implementation of the secondary task, which we used to sell our coverstory for the experiment and to focus participants' attention on the gender priming stimuli. It is possible that participants, in anticipation of the secondary task, delayed their responses to the primary one (e.g., to recall what the prime stimulus depicted, about which they would be asked a question soon). This could have attenuated priming-induced differences in reaction time.[14] Similarly to the concerns about our primary task, the best way of excluding such alternative explanations of our results would be to obtain evidence for non-gender-related sequential priming effects (e.g., "nurse" priming "hospital") using identical primes and an identical setup with these two consecutive tasks. If we can find evidence for such sequential priming effects but not for sequential gender priming effects, that would speak towards the difference between the priming effects (i.e., at the least, that gender stereotype priming effects are weaker than other semantic priming effects).

[14]If this was indeed a problem, it might not have impacted Karylowski et al. (2001)'s results. They do not report how much time passed between their primary and secondary task, so it could be a long interval (consistent with that, their interval of 2 seconds between prime and primary task is already unusually long). At long inter-task intervals participants might be inclined to prioritize tasks, one at a time – keeping the impact of priming on reaction times unattenuated.

Where to, now, priming? Our study is yet another demonstration that effects are fragile and designs become complicated, once one tries to collect evidence for stereotype priming while accounting for alternative explanations.

A question that Corneille and Gawronski (2024)'s paper "Self-reports are better measurement instruments than implicit measures" implicitly poses to proponents of implicit measures is what advantages implicit measures have over explicit ones. This is an important question to consider for any proposed application of implicit measures, given that they come with some meaningful downsides compared to explicit ones, like their generally lower reliability (Gawronski & De Houwer, 2014) and the difficulty of designing good implicit paradigms, which, besides being extra work, also limits what one can use them to test for (Corneille & Gawronski, 2024, e.g., it is much easier to ask people questions about parenthood than to receive answers about that via priming).

One potential advantage of implicit compared to explicit measures is that it might be easier to hide from participants what they assess, making potential negative effects from demand characteristics (Ruhl, 2023) less likely. Some version of this claim has commonly been made about implicit measures (see e.g., Cameron et al., 2012; Gawronski & Hahn, 2018), though, according to Corneille and Gawronski (2024) with little empirical evidence. We tried to collect empirical evidence of priming when participants are unaware of what we are trying to measure. This would have been a promising start, though still not a demonstration that implicit measures are superior in that quality compared to explicit ones (for which one too can invest efforts into disguising what they attempt to measure). Our findings are sobering in two respects: Not only did we not find evidence of stereotype priming, 11 of our 172 participants still correctly guessed the purpose of our study. That is despite us, compared to the average stereotype priming study, going through extreme lengths to disguise what we wished to measure (i.e., half of all trials of our main task and an entire secondary task per trial were included solely to sell our cover story).

Worryingly, the true number of cases that could perceive what the study is about could be higher still. Our 11 "hypothesis guessers" were all high performers relative to our included participants (i.e., none would have been excluded for performance-related criteria, and on average they were faster, $\delta_{RT} = -0.324$ sec, and more accurate, $\delta_{Acc} = 9.88\%$, than included participants). We see two different potential explanations for this relationship. Firstly, these participants might have simply been more intelligent than the average, non-excluded participant, which could explain both their higher performance and their higher likelihood of detecting the relationship between prime and target. Secondly and more worryingly, better task performance and them discovering what the study is about might have stemmed from these participants paying more attention to the task (e.g., because they were more motivated) than the average participant. This would imply that more participants would have discovered this relationship, had they taken the task fully seriously. As researchers generally want their partic-

ipants to pay full attention to a study and to perform it to the best of their abilities, that would be bad news. Follow-up studies could investigate this by manipulating participants' motivation to perform (e.g., by paying them bonuses based on their performance relative to others) and by testing whether that manipulation influences participants' chance of correctly guessing the hypothesis of the priming study.

More generally, researchers tempted to use implicit measures may be advised to 1) consider which property unique to implicit measures they believe to be invaluable for their intended application, 2) think about whether explicit measures could be modified to also have that property (e.g., Corneille & Gawronski, 2024, discuss how self-reports can be used to measure automatic processes), and if not, 3) empirically test whether the implicit measure actually possesses that desired property. We agree with Corneille and Gawronski (2024) that it is concerning for a field with decades of history to defend favorable assumptions about implicit measures with an absence of contradictory evidence rather than with the presence of empirical support. Here we wanted to provide such positive support for one assumption: that implicit measures are subtle about what they assess. We managed the "implicit" in that we could mislead most of our participants about the purpose of our experimental task. But we did not manage the "measure".

Chapter 6

Preamble – Why it is hard to evaluate the relationship between newspaper content and political opinions

"If you aren't careful, the newspapers will have you hating the people who are being oppressed and loving the people who are doing the oppressing."
— Malcolm X (1965)

The initial goal of our NWO project was to assess whether and how strongly newspaper content impacts its readers' (political) opinions. On some level the existence of such a causal relationship between content and opinions – as is assumed by the above quote by Malcolm X – is self-evident: Opinions are formed based on information people come across, and newspaper articles surely are one such source of information. But the relationship is hard to investigate empirically. Both experimental and observational approaches to such investigations have limitations.

Experiments, in which participants are assigned to particular media diets and one measures how that influences their opinions (see e.g., Coppock et al., 2018), are hard to design for at least two reasons: Firstly, media that is forcefully assigned might be consumed differently than media that participants consume organically (e.g., either sought out, spontaneously came across, or got recommended from trusted sources). For example, participants might be more hesitant to accept information forced on them. Consequently, the impact of experimentally assigned newspaper content might be different from the impact newspapers have in daily life (see e.g., Stroud et al., 2019, for an empirical demonstration of such differences).

Secondly, since people typically change their beliefs slowly (e.g., over the course of weeks or months, not minutes), short experiments likely are poor representations of how newspaper content impacts political opinions in daily life. But if experiments are conversely run over long time stretches, it is hard to ensure that people consume media as assigned (see e.g., Broockman & Kalla, 2022, who had to ensure through daily spontaneously-designed quizzes that participants watched the assigned TV channels) and harder still to measure what and how much additional, unassigned media participant consumed of their own volition. This in turn makes it difficult to evaluate the impact of the assigned newspaper information on people's opinions: What people choose to consume could impact their opinions more (e.g., because it outnumbers assigned sources in quantity), or it could interact in unforeseen ways with the information that people were assigned to consume (e.g., discrediting an assigned newspaper article, if the spontaneously consumed media presents a differing view on the same topic). Both these factors make it hard to experimentally estimate the extent to which newspaper consumption influences people's political opinions in daily life.

The alternative – naturalistically observing what people choose to read and relating that with their political opinions – circumvents these two problems, but is causally more ambiguous. If one finds a statistical relationship between newspaper content and their readers' political opinions, this could mean that newspaper content influenced people's opinions. But it could, conversely, also reflect readers seeking out publications that affirm their already-held political opinions (Newton & Brynin, 2001), and/or newspapers adapting their reporting towards (what they perceive to be) their core readership's preferences. In other words, it could be opinions that influence media consumption rather than the reverse. To some extent, this problem of causal directionality can be addressed by measuring opinions and newspaper consumption, repeatedly. If a change in media diet (i.e., in which newspapers are read, as well as in how much of them is read) precedes a change in political opinions, the change in political opinion cannot have lead to a change in media diet.[1] That being said, even this pattern is causally ambiguous: It could be that an unmeasured third variable first causes a change in media diet and subsequently a change in political opinions. For example, if a close friend changed their opinion on a political issue, one might at first be open to reading articles they recommend (i.e., change one's media diet) – while initially maintaining one's differing opinion – and subsequently succumb to the friend's arguments and adjust one's political opinion, as well. In other words: Even if media diet and political beliefs were independent, if media diet is easier to influence than beliefs, a third variable that influences both could lead to an aggregate pattern (across people) in which changes in media diet precede changes in beliefs.

[1] At least if one assumes that participants' expressed beliefs reflect their genuine internal beliefs. It could be the case that people's privately held beliefs began to change – which lead to a change in media that they sought out – but not yet enough to change which answer option they choose in a questionnaire.

Given these different theoretical challenges, we intended to develop a (causally suggestive) observational study step-by-step, iteratively addressing practical challenges and causal ambiguities, as they arise. The below-described dataset resulted from our first step in this process.

Why we collected the below-described pilot data

Until recently, there was an important practical challenge in addition to the just-described theoretical ones when designing naturalistic and observational studies of the relationship between newspaper content and readers' political opinions: When participants could freely choose which newspaper articles to read, the total number of articles across participants became too large to analyze in terms of content. Modern natural language processing (NLP) technology, which automates such linguistic analyses, makes this practically feasible, at least in principle (though at the potential cost of systematic biases in how such texts are evaluated; see e.g., Elbouanani et al., 2025). As a first step on our road to a (causally suggestive) observational study, we hence tried to find one or more NLP methods that measure the content of newspapers in ways that correlate with political opinions of the newspaper's readers.[2]

To that end, we collected two datasets in parallel: firstly, a dataset of questionnaire data – the dataset described in the data paper, below – on Dutch adults' demographics, newspaper diet, and political views. Secondly (not described here), a dataset of all publicly-accessible articles that were published in Dutch newspapers around the time we collected the questionnaire dataset, whose content we would analyze with NLP methods. Had we succeeded in finding one or more NLP methods whose scores for newspapers correlated with their readers' opinions (e.g., if an article was judged as opposing immigration, its readers also tend to oppose immigration), we would have collected a second sample of questionnaire responses and (concurrent) newspaper articles to validate said method(s) with new data from new participants.

We considered such a collection of new data necessary, for two reasons: Firstly, since our final observational study – in which we wanted to track how newspaper content *influences* readers' beliefs (i.e., requiring us to detect changes over time) – would have involved multiple measurement instances, our NLP method(s) had to be robust over time (e.g., not based on patterns in text that will quickly become outdated). Secondly, because we needed new data for out-of-sample validation of

[2]In our view, the absence of a relationship between newspapers' content and the beliefs of their readership is unlikely. No matter which of four plausible causal accounts were true – whether newspaper content influenced political opinions, political opinions influenced newspaper content, both influenced each other, or both were caused by a third variable – a positive correlation would exist. Consequently, we deem it unlikely that an NLP method picked up on politically relevant newspaper content, if we find no statistical evidence for such a correlation.

our chosen NLP method(s), to protect us from overfitting during model selection (i.e., inadvertently selecting a method, partly because of idiosyncrasies of the initial data – meaning that it generalizes poorly to new data). The latter was a serious concern, given that we planned on assessing several different NLP methods for analyzing the content of newspapers.

For our purpose of assessing newspaper content, we aimed to evaluate traditional NLP methods – like embedding-based bias measurement tools (e.g., Caliskan et al., 2017) and sentiment analysis methods (Jim et al., 2024, e.g., to see whether the sentiment in a newspaper article about immigration is predictive of its readerships' political preferences regarding immigration) – but wanted to also use the opportunity to evaluate a more recent collection of methods, called *synthetic sampling* or *silicon sampling*.

Silicon sampling In silicon sampling methods, large language models (LLMs) are prompted to emulate people and respond to questionnaires as those people would, based on information like the people's demographics, votes on previous political elections, responses to other questionnaire items, or social media posts (see e.g., Argyle et al., 2023; Bisbee et al., 2024; Cerina, 2025). Prompting approaches differ in how they specify the LLM's role. Some explicitly assign the model a persona and instruct it to answer from that person's perspective (e.g., "It is 2020. You are a 25-year-old single non-Hispanic white woman with a Bachelor's degree or more [...]. Provide responses from this person's perspective. Use only knowledge about politics that they would have."; Bisbee et al., 2024). Others implicitly provide the LLM with the to-be-emulated person's responses to previous questionnaire items (e.g., that measured their socio-demographics) and implicitly elicit target responses: For example, Argyle et al. (2023) delivers information in the form of previous answers to a mock interview and elicits the target response by leaving the LLM to generate the final answer of that interview (e.g., Argyle et al., 2023). Yet others combine such approaches. For example, Lee et al. (2024) used an interview-style prompt, but explicitly instructed the LLM to act as interviewee and to answer the final question. In most of these studies, attempts are made to predict variables related to politics (e.g., state polls of the US presidential election; Cerina, 2025; self-reported voting behavior; Argyle et al., 2023; Bisbee et al., 2024; von der Heyde et al., 2026; views of members of political in- and outgroups; Bisbee et al., 2024; opinions about climate change; Lee et al., 2024; i.a.). However, the same approach can be used to elicit LLM predictions for responses to any other questionnaire or interview question (e.g., how much people like different brands of soft-drinks; Kaiser et al., 2025). Generally, silicon sampling approaches are evaluated by how well they can reproduce the response patterns (i.e., generally means, and sometimes measues of spread like standard deviations) of different, often demographic, subgroups of people. The growing popularity of silicon sampling is exemplified by Gallup – one of the world's most

prominent public polling agencies – recently announcing plans to explore how such methods can be used to refine questionnaire items and hypotheses for their future surveys (Marlar & Ritter, 2026).

The broader project plan Provided that silicon sampling can successfully predict Dutch participants' political opinions, we could subsequently explore whether predictions improve when LLMs are also given information about participants' newspaper diet (e.g., by providing article headlines or full articles from newspapers that the participant reported reading). Improved predictions would suggest that LLMs extracted politically relevant signals from the newspapers. We could then use silicon sampling approaches for our main project: to test whether qualitative changes in newspaper diet (i.e., shifts in diet large enough to alter LLM's predictions) precede changes in participants' political opinions.

Such an application of silicon sampling approaches is conditional on these approaches successfully predicting Dutch participants' responses to (politically relevant) questionnaire items. To evaluate the extent to which they can do so, we aimed to address two questions: firstly, whether LLMs' predictions for people's responses are in fact predictive of people's responses. We expected demographic data to be predictive of political opinions (though see our results), so we reasoned that LLMs did not properly use the demographic data in their predictions for questionnaire items, if they cannot predict responses of participants better than chance.[3]

In a subsequent larger sample of participants (i.e., the second round of data collection that we wanted to use to validate our NLP methods), we could secondly evaluate to what extent LLMs' predicted responses for items of the same questionnaire are internally cohesive. For human responders, when different questionnaire items address the same underlying construct, we expect responses to them to correlate in predictable patterns. For example, if responses on one item indicate that the respondent is opposed to immigration (e.g., agreeing with the statement "We should stop all immigration to this country."), we expect that responses on other items are more likely to also indicate opposition to immigration (e.g., disagreeing with the statement "We need more immigrants in this country."). LLMs' predicted responses are cohesive to the extent to which their responses' correla-

[3]Had we found that LLM-predicted responses are predictive of human responses, we would have subsequently explored how well LLM predictions capture the distributions of human opinions. Previous research has found silicon samples to occasionally be directionally correct, but qualitatively different from human data: For example, Bisbee et al. (2024) found that responses from silicon samples had close-to-correct means for how people feel about racial groups but little variability, while human responses varied widely. And Kaiser et al. (2025) found that silicon samples could reproduce overall relative trends – that 4 famous compared to 4 more obscure soft-drink brands were more well-known and liked – but provided responses with too little variability and generally overestimated how well-known and liked the softdrinks were (e.g., for famous brands, the mean proportion of respondents who indicated knowing the brand was $\approx 90\%$ in the silicon sample, compared with below 65% among humans).

tional patterns resemble those of human respondents. We planned on testing this by assessing the degree to which measurement invariance holds across Structural Equation Models (SEMs) of human responses and LLM imitators' responses (for an optional and brief conceptual introduction to SEMs and measurement invariance, see Appendix C.1).

In principle, these two questions are orthogonal: It is possible that LLMs give internally consistent answers (e.g., answer all immigration questions similarly to how a person would who is opposed to immigration), but are incapable of emulating individual people (e.g., cannot, based on the sociodemographic information we provide about a person, predict whether the person is likely oppositional or supportive of immigration). However, since we were primarily interested in developing an NLP method for assessing prediction-relevant content of newspaper articles, the collection of a sufficiently large sample to address the second question was conditional on successfully developing a silicon sampling method that performed better than chance.

A conceptual overview of this dataset

Below, I will discuss the content of the piloting data we collected in exhaustive detail, but, to stick with the conventions of a "data paper", I will there not report on design decisions (why we include what we do), or on hypotheses we tested using the dataset. In the following two sections, I will hence touch on these points.

Main content The questionnaire – and, hence, the resulting dataset – contained questions about participants' socio-demographics (e.g., age, highest obtained educational degree, whether they had children, i.a.) and media diets (e.g., which newspapers they read, how much time they spend reading each, what other forms of media they consume, i.a.) which ultimately were meant to serve as predictors of political opinions.

European (Maxwell, 2019) and especially Dutch (De Vries, 2018) political opinions have been argued to largely fall on two dimensions: a left-right axis that addresses people's economic and (re)distributional beliefs and a cosmopolitan-parochial axis which addresses people's beliefs about international integration, international cooperation and migration. To tab these dimensions, we included a subscale from the 2014 version of the European Election Survey (EES; Schmitt et al., 2016) that assesses respondents' economic and redistributional beliefs and a subscale from the 2024 version of the European Social Survey (ESS; European Social Survey European Research Infrastructure, 2024) on immigration, as immigration is central to Dutch political debates (Harteveld, 2021).[4]

[4]We included this particular EES left-right scale, because De Vries (2018) used it in her analysis of Dutch voter opinions with which she introduced her two-dimensional (i.e., left-right, parochial-cosmopolitan) model of political opinions. We deviated from her choice for

Additionally, we included the "trust in institutions" subscale from the ESS which, on every question, simply asks respondents to indicate how much they trust a particular political institution (e.g., the Dutch police, the Dutch government, the European Union, i.a.) on an 11-point integer scale which was unlabeled, except for its two end points "0 – do not trust the institution at all" and "10 – complete trust". We considered it relevant to measure trust in political institutions, because such trust has been argued to be important for the stability and legitimacy of democracies (see e.g., Lühiste, 2006). The European Social Survey and European Election Studies are large-scale, repeatedly administered (i.e., ESS: every 2 years since 2002; EES: every 5 years since 1979) studies with established psychometric quality also for its Dutch translations (e.g., Zmerli & Newton, 2008), which is why we considered them high-quality measurement tools for our purpose.

We also included three subscales from Mansoor (2021) which, respectively, addressed how responsive a government was perceived to be, how transparent it was perceived to be, and how much respondents trusted it. In Mansoor (2021)'s SEM, participants' responses on the perceived responsiveness subscale (which Mansoor adapted from Vigoda-Gadot & Yuval, 2003) and perceived transparency subscale (adapted from Park & Blenkinsopp, 2011) were related to their responses on a "trust on [sic!] government" subscale (adapted from Grimmelikhuijsen, 2012), but there was no direct relationship between responsiveness and transparency. We included these scales to test whether LLMs, when imitating human respondents, could replicate such more complex patterns of relationships between questionnaire responses.[5]

Deception and distractor content We went through great lengths to conceal the purpose of our study, as we believed that asking participants about their media diet and political beliefs would make it apparent that we are investigating their association. Awareness of a study's hypothesis could influence participants'

the parochial dimension, though: De Vries included items from the EES that each addressed a singular different political topic, like immigration or EU integration. We instead focused on immigration and included several items for it (i.e., the ESS immigration subscale), because immigration was central to her second axis of opinions and additionally used by Harteveld (2021) as a proxy for his slightly differing second axis (i.e., both agree that the first axis is economic, Harteveld but not De Vries considers the second axis cultural).

[5]Mansoor (2021)'s structural equation model included several additional subscales, discrete questionnaire items, and relationships that we did not include. The model also included an additional indirect relationship between our 3 included subscales: perceived transparency and perceived responsiveness were related to a subscale of questions about how well the government responded to the COVID-19 pandemic, which in turn was related to responses on the "Trust on government" subscale. We only included the responsiveness, transparency and trust subscales, firstly because a simpler model (i.e., involving only 3 rather than 8 constructs) suffices as a proof of concept of LLMs' ability to replicate complex response patterns, and secondly because several of her model elements were much less relevant outside of her (i.e., COVID) research context.

behavior (e.g., people with extreme political opinions might endorse less extreme positions to prevent their favorite newspaper from being branded as extreme).

To that end we crafted a coverstory that would give participants a reason for why we asked them about their political opinions and media consumption that was unrelated to the relationship between the two – and the remaining items of the questionnaire (i.e., about two thirds of its total length) were distractor items to sell this coverstory. The coverstory worked reasonably well: Of 77 participants that ventured a guess about what we researched (for a chance to win up to 1 euro on top of the 5 euros they were paid for participating), only 11 provided guesses that were close to the truth, while 47 were too far off to earn any bonus.

As a framing device, we introduced our study by claiming that newspapers are "supposed to inform readers about topics that they find interesting and about topics that are so important that everybody should know about them" and reported investigating "a) whether Dutch newspapers [...] inform readers as they are supposed to, b) which people still regularly read newspapers, and c) which newspapers people in the Netherlands still read". Points b) and c) were supposed to provide cover for why we asked participants about their demographics and (media) consumption habits. To strengthen that cover, we included distractor questions that addressed attributes of participants unrelated to their sociodemographics (e.g., asked them how interested they were in different broad topics that newspapers report on), and distractor questions that addressed attributes of newspapers (e.g., asked them about the extent to which different newspapers are neutral in tone).

Point a), on the other hand, provided cover for our political questions: In 5 blocks themed by different topics that newspapers commonly report on – one for sports, entertainment, domestic events and domestic politics, international events and international politics, and science – we first asked participants a collection of questions about the topic and subsequently asked them to indicate whether different news events related to the topic had taken place and how certain they were that those did (not) take place. Half of the news items were fictitious, half pertained to events that had transpired in the two months before data collection. News items were chosen (or invented) to relate to the questionnaire items that occurred before. For example, to retrospectively justify the inclusion of the ESS's "trust in institution" scale in the "international events and international politics" section of the questionnaire, some of the news items that we asked about related to international institutions (see Appendix C.2).

The questionnaire items preceding the recurring "did this news event take place" part of each block were diverse in topics and response formats. The topics were diverse to obscure our hypotheses by suggesting multiple plausible research goals (e.g., some items, like "Hobbies were invented to distract us from what really matters", could give respondents the impression that we investigated conspiratorial beliefs). The response formats were diverse, so that the items relevant to our research question would not stick out (as e.g., the 6 questions of the ESS's

immigration subscale have 2 response formats and 4 sets of response labels; see Appendix C.3). Most of these questionnaire items were invented for this study and formed "subscales" of shared response formats, but whenever they fit with a block's topic and were potentially relevant, we instead included items from validated questionnaires. For example, we included the ESS's happiness item (i.e., "Taking all things together, how happy would you say you are? Use a scale from 0 - Extremely unhappy to 10 - Extremely happy") as a distractor item in the questionnaire section on participant characteristics. Similarly, we included the "Trust in Science" scale from Farias et al. (2013), which measures the extent to which respondents believe that science is a belief system that is epistemically superior and that can provide meaning to life, in the block on science news events.

We included the political items from the ESS and EES in the "news event" blocks on domestic and international politics among individual distractor items (e.g., about taxation, remilitarization, or the firmness of respondents' political beliefs). These blocks came 3rd and 4th, respectively, as we wanted to avoid that the hypothesis-relevant questions receive special attention that first or last blocks of questions might get. Appendix C.2 depicts the questionnaire block on international politics to provide examples for the diverse content and response formats that we included in our distractor items. For scales that we needed to translate – Mansoor (2021)'s 3 subscales and most items from the "Faith in Science" scale[6] – two of our researchers first independently translated all items and then resolved disagreements in a discussion. Machine-assisted back translations by a 3rd researcher confirmed that the final Dutch translations captured the content of the original English items.

Analyses and conclusions

Data and analysis scripts for all here-reported analyses are accessible on https://osf.io/23k68/overview?view_only=17da84b4c1074070980ad2069b4b20b7.

Preparatory steps Of our 105 participants, 11 were excluded for (coming close to) correctly guessing our hypothesis, and a further 4 were excluded for responding too quickly, as they likely did not properly complete the questionnaire (e.g., only skimmed the questions). For time exclusions we used a conservative criterion: Our estimated completion time for the questionnaire was 33 minutes

[6]Rutjens et al. (2018) compiled a short 5-item version that included 4 items from Farias et al. (2013) and one additional item "We believe too often in science, and not enough in feelings and faith" from the 1993 ISSP Environment module (ISSP Research Group, 1995; see also Hayes & Tariq, 2000). In Većkalov et al. (2023) they translated this 5-item version into Dutch. As we were unsure about which of the two versions has better measurement properties – Farkas's full set of items, or Rutjens's short version – we decided to include Veckalov's translations (including the ISSP item), as well as to translate the remaining Farkas items into Dutch.

(and the observed median $\approx$ 32) and we excluded participants that completed the questionnaire in less than 16 minutes. We collected data in a manner that ensured that the final post-exclusion sample included the same number of women as men.

Before we performed inferential analyses, we computed Cronbach's α values for all our included scales, as indices of their internal consistency. Responses to the "perceived responsiveness" scale, $\alpha = .571$, had low internal consistency (e.g., one that fell below the conventional .70 rule of thumb; though see Hussey et al., 2025 for a critique of such threshold practices), which is why we did not use this subscale in our inferential analyses. We deemed the internal consistency of all other subscales acceptable, $\alpha_{LeftRight} = .724$, $\alpha_{TransparencyGovernment} = .825$, $\alpha_{TrustGovernment} = .857$, $\alpha_{FaithInScience} = .884$, $\alpha_{TrustInstitutions} = .889$, $\alpha_{Immigration} = .910$.

As quick data quality sanity checks, we proceeded to a few simple correlation tests. With the exception of one – we found no evidence that participants' mean on the "faith in science" subscale was related to their educational attainment, Spearman's $\rho = .044$, $p = .68$ – all expected correlations were found: Higher educational attainment was associated with higher household income, Kendall's $\tau = .241$; mean responses on Mansoor's "trust on government" scale were associated with means on their "perceived [governmental] transparency" scale, Pearson's $r = .683$; with trust in politicians, as assessed by the item from the ESS "trust in institution" scale, $r = .573$; as well as with trust in "the current Dutch government" as assessed by an original item that we added to the "trust in institution" scale, $r = .415$ (all ps $\leq .002$).

(Exploratory) inferential analyses Silicon sampling approaches use socio-demographic information about participants to predict the participants' questionnaire responses (see e.g., Argyle et al., 2023; Bisbee et al., 2024; Cerina, 2025; Kaiser et al., 2025; Lee et al., 2024; von der Heyde et al., 2026, often alongside other information about participants like their political party affiliation). We planned on doing the same.

This is however only sensible if socio-demographic variables are in fact related to political opinions, within our sample. Consequently, we tested for such relationships with parallel series of frequentist linear mixed models, one for each of six (political) outcome variables: economic (left vs right) and redistributional preferences[7], views on immigration, trust in institutions, "trust on government",

[7]When we decided whether to proceed to the next project stage, we had not yet analyzed the EES's economic and redistributional "left vs right" scale, which we had initially collected to evaluate silicon sampling. One potential use was to test whether adequate prediction requires providing LLMs with direct information about people's political beliefs, thereby making the task partly circular (which is one potential way to read the results of Bisbee et al., 2024, and Lee et al., 2024). For another, we could use it to explore whether LLMs use newspaper content primarily to infer where readers lie on the left-right scale of political beliefs, or whether they pick

perceived transparency of the government, and faith in science.

In each series, we started with a complex model that contained as fixed-effect predictors participants' age, gender, household income and highest attained educational degree, as well as all interaction effects between these predictors.[8] Stepwise we reduced the complexity of these models until no further exclusion of a predictor was supported by the data. Specifically, we iteratively identified for the currently best model which predictors could be removed without rendering interaction effects uninterpretable (e.g., if a model included age and the age:gender interaction as predictors, only the latter could be removed), fitted all of these simpler models and identified a new best model based on the Bayesian Information Criterion (BIC). When no further simplification of models was supported, we similarly tested whether the inclusion of a predictor would be supported by the data.

We concluded that demographics were not predictive of our outcome variables: With the exception of immigration, the winning models included no fixed effect predictors (beyond the intercept). For immigration, too, the winning model was sparse: It included gender but no other predictors.

For ordinal variables, like our income bracket and obtained educational degree predictors, linear models estimate several different trends in the data (i.e., one fewer than the number of observed levels, so e.g., linear, quadratic and cubic trends for a factor with 4 levels) and hence involve many more degrees of freedom than do interval variables like age. As BIC penalizes models by their number of degrees of freedom, these two predictors might get particularly punished by our method of model selection (e.g., modeling the interaction between the main effects of education and income could lead to the estimation of up to 5x4 model terms). To test the robustness of our conclusions, we ran complementary model series with these two predictors (improperly) treated as interval level variables

up on more sophisticated patterns (i.e., we could test whether providing both newspaper content and left-right scale responses improves silicon sample predictions relative to only providing left-right scale responses). Since this subscale provides another set of political opinions, I decided, during the writeup of this chapter, to also model these beliefs. Results were substantively similar to those of other outcome variables, so initially omitting this scale did not impact the decision to discontinue the project.

[8]In all but one model series, the random effects were random intercepts for participant (i.e., accounting for the fact that participants tend to provide similar responses on items about the same topic) and random intercepts for test items (i.e., accounting for the fact that items systematically differ in how hard they are to endorse). In the model that predicted participants' views on immigration, we only included a random intercept for participants. Since half of the items for that subscale have 11 response categories and the other half have 4, we had to standardize responses within questions to make comparisons across questions meaningful (e.g., to make possible statements like "participants that were below average in their support of immigration on question 1 were also below average in their support on question 4"). That made modeling differences between these questions statistically meaningless – all items, post standardization, had the same mean and standard deviation.

(i.e., in essence, we tasked the statistical model to only capture linear trends in the data, and to ignore quadratic and other higher-order trends). This did not change our results; the winning models were identical.

Takeaways for our project In this study, we specifically assessed the impact of three socio-demographic variables that had in a meta analysis been found to be individual-level predictors of, at least, views on immigration (Dražanová et al., 2024): age, education, and income – the latter two of which are core components of socioeconomic status (Galobardes et al., 2006), which, too, is thought to be robustly predictive of, at least, views on immigration (Maxwell, 2019). Additionally, we picked gender, as it is a frequently measured demographic marker and one that is generally believed to be predictive of different political preferences (see e.g., Dassonneville, 2021; Shorrocks, 2018).

We were surprised to find that these demographics were largely unpredictive of participants' political opinions (and faith in science) in our sample. Across our six outcome variables, we found only one effect: While Dražanová et al. (2024)'s meta analysis found no evidence for a robust effect of gender on such views, women in our sample were on average about half a standard deviation more positive about immigrants' contributions to society and more supportive of open immigration policies than men, $\beta_{Gender} = 0.497$.

With this data, we wanted to evaluate the extent to which LLMs can emulate human respondents based on their demographics (i.e., to evaluate silicon sampling approaches). Had we found convincing evidence of LLMs' ability to emulate human respondents, we could then proceed to explore whether and how information about participants' media diet (e.g., providing LLMs with articles from newspapers that participants reported reading) could aid LLMs in predicting people's responses. Given the results of our exploratory analyses, we could not meaningfully evaluate silicon sampling approaches with our data: Failure of LLMs to provide above-chance predictions for our participants' responses (based on their demographics) could reflect the absence of a relationship between demographics and responses in our sample rather than LLMs' inability to emulate humans.

We hence did not proceed to the next stages of this project, in which we would have first developed silicon sampling approaches (i.e., replicated influential past approaches – like Argyle et al.'s, 2023, and Bisbee et al.'s, 2024 – and tested our own variants) and subsequently collected data from a new and larger sample of participants to evaluate these approaches on new data and to assess the measurement invariance between human and LLM imitators' responses. In the data paper below, we report on the piloting dataset, so others may use it for their own research (for a discussion of different potential use cases, see Section 6.4).

Limitations In retrospect, there are two design choices we would have made differently. Firstly, since participants had to opt in to guess the study's purpose, we could not assess whether the 26 who did not opt in had perceived it. If some of them guessed correctly and altered their responses (e.g., gave more socially desirable responses to political questions, if they knew that those were our research focus), that might have attenuated (or otherwise distorted) the relationship between (socio-)demographics and political beliefs in our data. While we deem it unlikely that this significantly influenced our results – since participants were financially incentivized to guess correctly, we suspect that most that had a clear idea did guess – a better choice would have been to display this question by default. This is how we later implemented that question in the priming study (Chapter 5), where only 1 of our 100 final participants opted out of guessing.

Secondly, we did not include explicit attention checks (e.g., questions like "To indicate that you are still reading the questions, please click on answer option 2."). Consequently, some participants that provide random or inattentive responses to many items, probably went undiagnosed. This could similarly attenuate the statistical power we had to detect relationships between (socio-)demographics and political beliefs, compared to collecting data from additional participants (as we e.g., did for participants that we excluded for guessing our hypothesis). Initially we planned on using participants' awareness of particularly "unavoidable" news events as attention checks (e.g., exclude participants that disagree that Donald Trump won the 2024 US presidential election), however we ultimately decided against applying them, as it would be impossible to distinguish people that did not pay attention from people that reported on genuine but false beliefs.

Fitting for a dissertation about researchers' implicit assumptions (e.g., that "larger" necessarily means "better" for benchmark datasets; Chapter 3; or that evidence for stereotype priming can be obtained when participants are unaware of what is being studied; Chapter 5), the biggest flaw of this project was born of one such faulty assumption. As associations between (socio-)demographics and at least views on immigration seemed robust (based on the meta analysis of Dražanová et al., 2024), and given that political organizations act as though (socio-)demographics matter (e.g., women and "young people" are the most and 5th most appealed to groups in UK party manifestos, respectively; Thau, 2019), we incorrectly assumed that these associations must be large and easy to detect. In retrospect the flaw in that implicit assumption is obvious – what is true in large numbers of people (e.g., as are analyzed in meta analyses, and as are appealed to by political actors) does not necessarily have to be true for modest sample sizes – but while planning the project we did not even consider the possibility that we would find no such association in the piloting data for our six outcome variables.

Even though this was nominally a pilot study, it might have been wiser to collect a large sample of data (e.g., 400 participants, post exclusions). For one, this would have made results less ambiguous in regards to whether we simply

lacked statistical power to detect (real but small) associations between socio-demographic variables and political beliefs, or whether such relationships genuinely do not exist in the population we tested. For another, even if we had still found reasons not to proceed to the next stages of our project (i.e., to collecting a second sample of data to validate our NLP methods), the below-described dataset would have been more valuable. Since we have data for only a modest number of respondents, the properties of our many new questionnaire items (should researchers be interested in adopting them), and of our new Dutch translations for existing scientific subscales cannot be evaluated with sophisticated psychometric models (e.g., structural equation models). That being said, we prioritized creating a multifaceted dataset – one that can be used to explore a lot of different, interesting questions (e.g., whether newspapers can be roughly placed in the 2-dimensional space of Dutch political opinions by De Vries, 2018; how respondents perceive the impact of different scientific disciplines on their life; how choice of newspapers and/or time spend reading relates to expressed political opinions; i.a.). Consequently, we still found it worth sharing.

Data on newspaper consumption, political opinions, awareness of news events, and trust in science for Dutch adults in 2025

Dominik Bachmann[1,2], Oskar van der Wal[1], Maximilian Spliethöver[3], Willem Zuidema[1], Leendert van Maanen[2], Katrin Schulz[1]

[1] Institute for Logic, Language and Computation, University of Amsterdam
[2] Department of Experimental Psychology, Utrecht University
[3] Institute of Artificial Intelligence, Leibniz University Hannover

Abstract

We present a dataset on Dutch adults' views on newspapers, their political opinions, awareness of news events, and trust in science. Data were collected in early 2025 in a cross-sectional online survey on Prolific ($N = 105$; 54 male, 50 female; $M_{\text{age}} = 32.5$). The survey covered participants' socio-demographics, media diet (e.g., which newspapers they read, and how much), views on different Dutch newspapers (e.g., how factual, or biased they appear) and political opinions (e.g., views on immigration, redistributional beliefs, how much they trust different political institutions, i.a.). Additionally, the survey assessed participants' awareness of factual and fictitious newsworthy events (most of which from November and December 2024), as well as their other opinions on topics related to five different domains that newspapers report on: sports, entertainment, domestic news and politics, international news and politics, and science. Heavily annotated data (to enable immediate use) as well as the questionnaire used to generate the data are available on OSF. Researchers may reuse this dataset to investigate the interplay between newspaper consumption, socio-demographics and readers' opinions on a host of different topics.

6.1 Background

With this paper, we present a dataset about Dutch newspapers and their readers. It was primarily designed to test the association between newspaper consumption, readers' sociodemographic attributes, and their opinions. Secondarily, it was designed to assess newspapers – for example in regards to the extent to which they inform people about events that occurred.

Newspapers and (democratic) politics Newspapers have the potential to influence how democracies function (Hanitzsch & Vos, 2018): They can tell readers which events and political issues are worth paying attention to – "setting" political agendas (Entman, 2007). They can frame these issues and events – influencing readers' understanding thereof (Entman, 2007; Hanitzsch & Vos, 2018). How newspapers represent different groups (Brewer et al., 2017; Djourelova, 2023; Keita et al., 2024) can influence how readers perceive them. Newspapers can monitor powerful people, like politicians, for example to hold them accountable (Hanitzsch & Vos, 2018; Kalogeropoulos et al., 2024).[9]

Given these many ways in which newspapers can impact democratic functioning, a lot of scholarship on the roles of newspapers and many experiments on opinion formation through media consumption (e.g., Brewer et al., 2017; Coppock et al., 2018; Stroud et al., 2019) focused on the (complex, i.e., presumably at least bidirectional; Newton & Brynin, 2001) relationship between newspapers and political opinions. Our dataset, too, includes several questions about the political opinions of readers and apparent political positions of the newspapers they read. In our choices of what political content to investigate, we were partially influenced by the theoretical framework of De Vries (2018), who categorized Dutch political opinions along two dimensions: a economic and redistributional "left versus right" dimension, and a "parochial versus cosmopolitan" dimension of political focus. Among our questions on political opinions, we hence included some that could be used to place readers and newspapers on these two dimensions.

In addition to these questions related to politics, we also included questions that can be used to evaluate newspapers in regards to their non-political functions, like their role as informers (Hanitzsch & Vos, 2018), entertainers (Hanusch, 2019), or archives of text that future people can refer back to, to understand how people of a given era thought (Ringel, 2023). Specifically, we included questions that assess how well readers are informed about news events and their opinions on several topics (e.g., sport) that are commonly addressed by "lifestyle journalists" (Hanusch, 2019).

6.2 Methods

6.2.1 Study design

The data were collected with a cross-sectional online survey that was hosted on Qualtrics. Participants stemmed from the recruitment website Prolific

[9] In addition to such roles in democratic societies, Hanitzsch and Vos (2018) also lists several political functions that newspapers can fulfill in non- or less democratic societies (e.g., as a mouthpiece of the government, or to indoctrinate). We did not collect data on newspapers for these roles, as the Netherlands is a democratic nation in which several of these alternative roles (e.g., as a "mouthpiece" of a uniform powerful government) hence do not apply.

(https://www.prolific.com/). After providing their informed consent, participants were screened out if they reported living in a country other than the Netherlands, as it was part of the advertised eligibility criteria, alongside adult age, fluency in Dutch and an approval rate on Prolific of above 98% (i.e., the percentage of times past researchers accepted a participant's work, as opposed to refusing to pay them; e.g., because they did not correctly perform a study).

Eligible participants answered questions about themselves (including but not limited to questions about their sociodemographics and living situation), questions about their media diet (with a particular emphasis on their newspaper diet), questions about famous Dutch newspapers as well as whichever newspapers they reported reading, as well as five blocks of questionnaire subscales grouped by different topics that newspapers frequently report on: sports, entertainment, domestic events and politics, international events and politics, as well as science. At the end of the experiment, participants had the option to guess the purpose of our study as well as to receive a debriefing.

6.2.2 Time of data collection

Data were collected from January 30th until (and including) February 4th of 2025.

6.2.3 Location of data collection

All participants self-reported to be fluent in Dutch and residents of the Netherlands, at the time of data collection. As agreed upon with our ethics committee, we deleted any more detailed location (e.g., IP) data, immediately following data collection, so we cannot provide more detailed information of participants' current location at the time of data collection. That being said, one of the questions from the questionnaire section about personal information asked participants to indicate in what kind of a settlement they lived from "Small village (less than 3000 people)" up to "Big city (more than 300.000)".

6.2.4 Sampling, sample and data collection

In total, 105 participants completed the questionnaire, of which we excluded 4 for providing responses that we consider too fast (i.e., in below 16 minutes, given the median completion time of 32 minutes) and, after manual review of their answers, 11 for guessing (too close to) our tested hypothesis about the relationship between newspaper content and political opinions. We sampled in waves, to ensure that we had an equal number of male and female respondents (i.e., 45-45 of our 90), post exclusions. This resulted in our last 23 participants being female, as we at first had a gender imbalance. As users of our data may make different exclusion choices (e.g., retain participants that guessed our particular hypothesis, or exclude

Table 6.1: Participants' highest attained educational degree

Education attained	N (%)
Primary school	0
Highschool (VMBO, HAVO, VWO)	13 (12.4%)
Master craftsperson (i.e., "Meester")	0
Secondary vocational education (MBO)	13 (12.4%)
Bachelor's degree (HBO, WO)	47 (44.8%)
Master's degree (HBO, WO)	30 (28.6%)
Doctorate (i.e., PhD or MD)	2 (1.9%)

participants more aggressively based on completion time), we will below report the demographics for the full dataset.

In our full 105 participant sample, 54 reported their gender as male, 50 as female and one preferred not to answer the question. Participants had a mean age of 32.5 (SD = 10.5) and were quite highly educated (i.e., 69 obtained a Bachelor degree or higher; see Table 6.1). Most participants were employed, with only 11 being unemployed and 20 reporting to currently be students. Participants were paid 5 euros for participating, and had the chance to win up to an additional 1 euro for correctly guessing what we intended to research.

6.2.5 Materials / Survey instruments

Here we provide a conceptual overview of the questionnaire that generated this dataset. We shortly describe the measured constructs and groupings of questions within our questionnaire, but leave out details about individual response labels and, whenever a question is applied to different entities about which exact entities we asked about. For details on possible response categories and on different evaluated entities, see Appendix C.3. While we do not list the original Dutch formulations of questions or response categories in that Appendix, those are saved in the csv file that contains this dataset.

Following an initial screening (e.g., to screen out participants that were not residents of the Netherlands) and informed consent section – in which we also asked participants about their age and gender – the questionnaire had three sections with different foci.

Media In the first section of the questionnaire, we asked participants about their media diet and their opinions on newspapers that they have read. First we created an overview of where participants received their news from: We started by asking participants how many (fractions of) hours of the previous week they had spent reading online or print newspapers. Subsequently, we asked whether

they had in that week read one or more of a list of famous national and regional newspapers, and allowed them to list newspapers they had read which we did not list. Next we asked participant to estimate how their last week's reading time was distributed across the newspapers they reported reading (i.e., they had to allocate integer percentage points across all these newspapers which added to 100). Finally, we asked participants to choose one or more sources from which they receive (at least part of) their news – TV, radio, friends or family, newspapers, social media, or "other" (which they subsequently specified) – and subsequently asked them to estimate what percentage of their news they received from each of these different sources.

After thus obtaining an overview of their media consumption, we asked participants to evaluate different newspapers. Each participant evaluated all newspapers that they reported having read the week before, as well as each of the four (at the time) most read national newspapers in the Netherlands (i.e., De Telegraaf, Algemeen Dagblad, De Volkskrant, and NRC; Nationaal Media Onderzoek, 2025) that they reported knowing well enough to evaluate. With individual questions that all had symmetrical Likert style labels, we asked participants to evaluate newspapers in regards to: 1) the quality of their reporting, 2) the extent to which they are politically biased, 3) the extent to which their reporting is accurate and factual, 4) the extent to which they focus on regional versus international news, and 5) the extent to which their reporting is sensationalist versus neutral in tone. Finally, setting the stage for the third section of the questionnaire (in which questions were grouped by topic that newspapers report on), we asked participants about their personal interest in 18 different topics that newspapers report on.

Participants With the second section of the questionnaire, we collected sociodemographic data about our participants. First, we asked them about where they were born and in what job sector, if any, they work in. Subsequently, we asked them about their living situation: Individual Likert style questions pertained to their relationship status, to whether they have children, to their house hold income, and to where they reside (i.e., which of 7 answer options, including "small village (less than 3,000 people)" and "Large city (more than 300,000 people)", best describes the place where they live). The final Likert style item of this section asked participants for the highest educational degree they obtained. Subsequently, we asked participants to indicate their political orientation and happiness on two largely unlabeled 11-point scales with the respective anchor labels of "0 – left" and "10 – right", as well as "0 – Extremely unhappy" and "10 – Extremely happy". Finally, we asked participants to describe themselves "as a person" in 5 to 10 sentences of open text.

Newspaper content The last section of the questionnaire had individual blocks of questions that were grouped by domains that newspaper news items often re-

port on: sports, entertainment, domestic events and domestic politics, international events and international politics, and science. Each block concluded with a list of news events – partly fictitious, partly factual; all related to a block's respective topic (see Appendix C.4 for an overview of events by topic, as well as their veracity) – and participants had to indicate whether that event had taken place on a response scale of "I am certain this happened", "I think this happened but I am not certain", "I don't think this happened but I am not certain", and "I am certain this has not happened".

Besides these recurring and block-concluding event questions, each topic block included different questionnaire items. In the sport block, participants first had to indicate to what extent different competitive activities, like chess, or car racing, constitute "real sports". Subsequently, we measured their agreement with different individual statements that related to amateur and professional sports but were largely unrelated to each other (e.g., whether "for good health, exercising is more important than healthy eating", or whether it is "less valuable" to practice team sports instead of individual sports; see Appendix C.3). Next, we asked participants to what extent they trust different stakeholders, like amateur athletes or voting members of sports clubs, to advocate for changes that are healthy for the sport.

In the entertainment block, we similarly asked participants about their agreement with different arts- and entertainment-related questions that spanned different topics (e.g., whether TV stations have become obsolete, or whether poetry is overrated). Additionally we asked them about their perception of motives that people with certain artist and entertainer occupations (e.g., actors, orchestra musicians) usually have; whether people with these occupations are primarily interested in benefiting themselves or in bringing other people joy.

In the block on domestic events and domestic politics, we first asked participants 6 questions about their views on immigration policies and on immigrants (European Social Survey European Research Infrastructure, 2024): the first three questions pertained to their support or opposition for different people (i.e., of the same racial or ethnic group as most Dutch people, of a different racial or ethnic group, and of people from poorer countries outside of Europe) to immigrate to the Netherlands. The last three pertained to their perception of the impact that immigrants have on the Netherlands' economy, cultural life, and quality as a place to live. Subsequently, a singular item asked them about their perception of the tax burden in the Netherlands (i.e., whether it is too low, just right, or too high). Then we asked participants three questions about their economic and redistributional beliefs (Schmitt et al., 2016): the extent to which they favor or oppose state intervention in the economy, redistributing wealth from the rich to the poor, and raising taxes to expand public services.

Next were two individual items that measured the perceived frequency with which Dutch politicians speak to their voters (with options like "too rarely", or "too often"), and how urgent they think it is for the Netherlands to recruit more

soldiers. Finally, rounding up this block, we asked participants about the extent to which they agree with different statements, that all pertained to the extent to which they trust the government (Grimmelikhuijsen, 2012), and perceive it as transparent (Park & Blenkinsopp, 2011) and responsive (Vigoda-Gadot & Yuval, 2003).

The block on international events and international politics began with an open question. Respondents were asked to finish the sentence "Over the last 10 years, politics in Europe has...". Subsequently, we asked them to indicate the extent to which their current political beliefs are flexible or firmly held. Then we asked participants to indicate the extent to which they trust different institutions that were either explicitly political (e.g., the United Nations) or related to politics and the state (e.g., the police, or the legal system). Finally, we asked them in two separate clusters of questions to indicate for a list of countries that is either globally important (e.g., China) or has historic ties to the Netherlands (e.g., Indonesia) whether they were interested in news about said country and how they expected that country to develop economically (i.e., whether it will, in the next decade, become more or less prosperous).

We opened the final block, on science, with an item on whether they personally know a scientist (or are one themselves), in case that is predictive of their responses on the following clusters of questions. Subsequently, we assessed the extent to which respondents believe that science is an epistemically superior belief system and that science can provide meaning to life (Farias et al., 2013; Rutjens et al., 2018). Subsequently, we asked about their endorsement of individual science-related statements, most of which addressed the place they see science occupying in society (e.g., "Scientists almost never look at topics that could improve my life" and "Ordinary citizens should be able to vote on which topics research is conducted on"). Finally, we asked them in two separate clusters of questions to indicate for a list of academic disciplines how the discipline impacted their life (i.e., whether it on aggregate improved, worsened, had a neutral impact, or had virtually no impact on their life), and whether or not it is a "real science".

After this third and longest section of the questionnaire, we asked participants whether they wanted to guess what we attempted to research with this questionnaire (i.e., participants who agreed to guess could then leave an open text response) for a chance of winning a bonus payment, and whether they wanted to be debriefed. Subsequently, the questionnaire concluded.

Item originality The majority of questions in this questionnaire were specifically formulated for it. That being said, whenever relevant, we included previously validated questionnaire items.

In the first section of the questionnaire, we included one such item: The most recent round of the repeatedly administered European Social Survey (ESS; European Social Survey European Research Infrastructure, 2024) includes an item

that asks participants about their interest in politics. We adopted their question formulation and response options, but instead of asking about politics only, asked also about other domains of interest, such as sport and boardgames.

The question on participants' subjective happiness, in the second section of our questionnaire, was fully adopted from the ESS. Additionally, we administered a scale from the European Election Study (Schmitt et al., 2016) that assesses participants' economic leftwing versus rightwing and distributional beliefs. The item that simply asked participants to classify themselves as left- or rightwing was included in the second section of the questionnaire, as we thought it fit well with the section's other questions on self-identity and demographics. The remaining three items of that subscale instead pertained to economic policy preferences (i.e., their views on state intervention in the economy, redistribution, and taxation to fund public services) and were hence presented in the block on domestic politics in the third section of the questionnaire.

That same block also included 6 items from the ESS about participants' policy preferences on immigration (e.g., whether they want the Netherlands to allow in immigrants from poorer countries outside of Europe) and their views on immigrants (e.g., whether they undermine or enrich cultural life in the Netherlands). Additionally, this block included items from scales that measure how much participants trust their government (Grimmelikhuijsen, 2012), and how transparent (Park & Blenkinsopp, 2011), and how responsive (Vigoda-Gadot & Yuval, 2003) they perceive their government to be. These questionnaire items were previously administered with one shared response scale (Mansoor, 2021), a practice we adopted here.

In the block on international events and international politics, we included a scale from the ESS that measured how much participants trusted different (political) institutions. We used their response format, question formulation, and institutions, but also included two additional ones: "the current Dutch government" and "Dutch newspapers". In the block on science, we included Farias et al. (2013)'s "Faith in Science" scale, as well as the item "We believe too much in science and not enough in feelings and faith" from the International Social Survey Program (ISSP Research Group, 1995), as Rutjens et al. (2018) introduced a short 5-item "Faith in Science" scale, which includes this item alongside 4 items from Farias et al.'s scale.

We adopted previously-used Dutch translations of items, wherever possible. Translations existed for all items from the European Social Survey and the European Elections Study, as well as for the 5 items of the shortened "Faith in Science" scale (i.e., first introduced by Rutjens et al., 2018; subsequently translated in Većkalov et al., 2023). For other items from pre-existing questionnaires (e.g., the remaining "Faith in Science" items), two native-speaking researchers (OvdW and LvM) first independently translated all items from English to Dutch and then resolved disagreements in a discussion. Machine-assisted back translations by another researcher (DB) confirmed that the final Dutch translations captured

the content of the original English items. For all remaining items, one researcher (OvdW) translated items, and another (DB) double-checked translations.

6.2.6 Quality Control

On the individual level, users of this dataset may choose to exclude participants based on their completion time (i.e., in light of the intended and actual median length of 32 minutes), and based on their responses to purported news items (e.g., if they report being certain that Johan Cruijff – a celebrity with high name recognition in the Netherlands who died 9 years prior to data collection – made a comment last month).

On the aggregate level, we found statistical evidence for several common-sense correlational patterns (e.g., between attained education and household income, and between responses on our individual trust in "the current Dutch government" item and participants' mean response on Grimmelikhuijsen's scale of the same construct, 2012) that we performed as basic sanity and data quality checks.

We assessed the internal consistency of all pre-existing scales (see "Item originality"). With the exception of Vigoda-Gadot and Yuval (2003)'s "perceived responsiveness" scale, $\alpha = .571$, all had acceptable-to-good values for Cronbach's alpha, $\alpha_{LeftRight} = .724$, $\alpha_{TransparencyGovernment} = .825$, $\alpha_{TrustGovernment} = .857$, $\alpha_{FaithInScience} = .884$, $\alpha_{TrustInstitutions} = .889$, $\alpha_{Immigration} = .910$.

6.2.7 Data anonymisation and ethical issues

We obtained ethical approval for the collection of this data by the Ethics review board of the Faculty of Science at the University of Amsterdam.

All participants provided informed consent (prior to answering questionnaire items) by reviewing and agreeing to the project's consent form on Qualtrics. Consent also covered the publication of their (anonymized) data (see column "Consent.form" in the dataset for the exact phrasing of the consent form).

Potentially identifying information that is automatically collected by Qualtrics (i.e., each participant's Prolific ID number, their IP address, as well as their latitude and longitudes when they submitted their response) were deleted from the dataset before its publication.

6.3 Dataset description and access

6.3.1 Repository location

The dataset is accessible at https://osf.io/23k68/overview?view_only=17da84b4c1074070980ad2069b4b20b7.

6.3.2 Object/file name

The file "Full_Anotated_Dataset.csv" includes the dataset, "Qualtrics_Visual.pdf" a visual printout of the questionnaire's appearance, "Qualtrics_Questionnaire.qsf" the questionnaire itself, "Prolific_Ad.html" the text we used to recruit participants, and "Read.Rmd" the script we used for preprocessing and annotating.

6.3.3 Data type

Supplementary files Alongside the dataset itself – which is in csv form – the repository includes supplementary files: Enabling (direct) replications, the .qsf file can be used to import our questionnaire into Qualtrics. The .pdf printout provides an impression of what the questionnaire approximately looked like to participants (though the printout is in English, while participants saw a Dutch version of this questionnaire).

The last two files are included for transparency's sake, but are not required to use our data: The .html file includes the exact text we used to recruit participants on Prolific (alongside an English translation). The .Rmd file includes the script that we used to preprocess and annotate our data. This script cannot be fully run by users of our dataset, as we do not wish to provide access to non-anonymized data.

Dataset The dataset itself contains primary data that is only lightly preprocessed: Firstly, we excluded potentially identifying information (see Section 6.2.7). Secondly, we included two additional columns that reflect choices we made as researchers which other users of the dataset may choose to utilize in their analyses: a column that indicates how much of a bonus we paid participants (i.e., a proxy for our subjective evaluation of how close they were to guessing the particular hypothesis we tried to research – the relationship between which newspapers participants read and the political opinions they hold), and a column that indicates whether we excluded a particular participant from our analysis (i.e., either for responding too quickly or for coming too close to guessing the hypothesis of our study).

Finally we included 4 additional rows of text that describe the column of data to facilitate usage of our dataset: Row 3 of our dataset includes the exact Dutch formulation of the question that participants were asked which led to the generation of the data in this column. Row 1 includes an English translation of that question. Rows 2 and 4, respectively, provide an English description of all values that this column's question can generate (e.g., a description like "[integer; minimum 16, maximum 120]", or a mapping from numeric values to their Likert labels), and Dutch translations of values' labels, whenever those occurred in the questionnaire.

6.3.4 Format names and versions

The dataset is saved in long-format (i.e., one row per participant, one column per data point) in a .csv file. We include supplementary files of other formats in our repository, but these are optional (e.g., a visual indication of what the questionnaire approximately looked like to participants): Since the first four rows of the dataset include a codebook for all obtainable values (per question) and the exact question phrasings that we utilized, alongside their English translations, the dataset can be used on its own with any program that can operate on csv files.

6.3.5 Language

The data is stored in American English, though the exact Dutch phrasings of questions (and response options) are included as well.

A few of the questions asked participants to provide open text responses, for those we include participants' original Dutch responses, rather than English translations thereof.

6.3.6 License

CC-By Attribution 4.0 International

6.3.7 Limits to sharing

There are no limits to sharing.

6.3.8 Publication date

The dataset was published on OSF on August 17th 2026.

6.3.9 FAIR data/Codebook

Findable To make our project more findable, we included descriptive metadata in our OSF repository: a description of the project, the subjects it covered, and tags.

The repository includes a README file (that briefly explains the content of the files in the repository) which aids with navigating through it.

Accessible On OSF, data are easily accessible, as they can be accessed and downloaded without authorization (or similar) procedures.

Interoperable The dataset is saved in a csv file – a file format that is trivial to operate on for all (at the time of writing) commonly-used data analysis programs. The file itself contains a codebook of the exact questions that participants were asked, of the possible values their responses could take on, and of any other information (e.g., which news events were fictitious vs real) that might be required to analyze the data.

In summary: The csv file on its own should suffice to use this dataset.

Reusable The study materials, including the consent form, advertisement on Prolific (with which we recruited participants) and the questionnaire files are all included in the online repository. Direct replication of our questionnaire study would be possible with the dataset (i.e., csv file) alone – and is further aided by the supplementary files in our repository (e.g., they include the exact Qualtrics file we used to collect responses).

6.4 Reuse potential

6.4.1 Secondary analyses

In this PhD dissertation chapter, statistical models are reported that sought to predict participants' responses to 6 scales – the scales assessing "faith in science" (Farias et al., 2013; Rutjens et al., 2018), views on immigration (European Social Survey European Research Infrastructure, 2024), economic and redistributional beliefs (Schmitt et al., 2016), trust in different institutions (European Social Survey European Research Infrastructure, 2024), perceived transparency of the government (Park & Blenkinsopp, 2011), and trust in the government (Grimmelikhuijsen, 2012) – based on four sociodemographic factors (i.e., age, gender, household income, and attained educational degree). No other analyses have been performed with this dataset. This leaves many potential secondary analyses, of which we highlight three broad topics.

Policy preferences For one, researchers interested in the relationship between socio-demographic factors and political opinions (e.g., on immigration Dražanová et al., 2024) could select different socio-demographic variables than the four we analyzed. For another, researchers interested in political opinions could investigate the relationships between the different political opinions we investigated here (e.g., whether participants' redistributional beliefs are related to their preferences for immigration policies; or how predictive participants' self-identification as "left" or "right" actually is of their economic and redistributional beliefs, see e.g., Bauer et al., 2017). Similarly, researchers could investigate the relationship between participants' views on a topic (e.g., views of science as a meaning-providing

belief system) and their policy preferences related to the topic (e.g., whether they wish that citizens got to vote on which topics are researched).

Newspapers Researchers interested in newspaper consumption in the Netherlands – which newspapers are read, how much time participants spend reading newspapers, which personal attributes are predictive of people being newspaper readers, how Dutch newspapers are perceived (e.g., in regards to sensationalism, factuality, political bias, i.a.) – will find much to analyze in our data. Similarly, many analyses could be performed to investigate the relationship between newspaper consumption and political positions of their readers: for example, whether participants that read newspapers that others judged as "biased in favor of the political right" are more likely to endorse positions that are associated with the political right. Following De Vries (2018)'s two-dimensional model of political opinions, we not only included questions that tab the traditional left-right economic axis (i.e., for people: the EES scale; for newspapers: the question on which newspapers are politically biased in favor of the left or right), but also items that proxy the parochial-cosmopolitan dimension (i.e., for people: views on immigration, as Harteveld, 2021 and De Vries, 2018 highlighted that topic as a particularly important proxy; for newspaper: the question on whether a newspaper focuses too much on local or international events).

News events Researchers interested in the quality of newspapers (as providers of information) and researchers interested in misinformation could investigate which attributes of participants (e.g., sociodemographic characteristics, expressed interest in a particular domain of news) and of newspapers (e.g., hours read per week) are predictive of participants correctly identifying news events as factual or fictitious. Researchers interested in misinformation might similarly explore which attributes of our fictitious news events made them more or less believable (or conversely: why certain factual news events were particularly often misclassified as fictitious). Researchers interested in metacognition (Fleming, 2024) could investigate to what extent participants' subjective certainty of a classification is predictive of them correctly identifying a news event as factual or fictitious.

6.4.2 Psychometrics

Our dataset includes novel Dutch translations, that researchers may choose to adopt, for 4 scales that have repeatedly been used in past research (i.e., for Grimmelikhuijsen, 2012; Park & Blenkinsopp, 2011; Vigoda-Gadot & Yuval, 2003, and for most items of Farias et al., 2013). Additionally, we introduced several novel questionnaire items in Dutch and English. Researchers may choose to preliminarily explore their psychometric properties, if they are interested in adopting or adapting these questionnaire items. Due to the breadth of constructs that

we assessed, these researchers can also make preliminary judgments about the convergent and divergent construct validity of these items (e.g., evaluate whether the item "Science can pose a danger to human existence" assesses a construct distinct from the one assessed with the "faith in science" scale).

In addition to scholarly work, the dataset is also sufficiently multifaceted to be didactically useful: Students in statistics and introductory psychometrics courses can formulate a host of different hypotheses, given our rich collection of (socio-)demographic factors and questionnaire responses and subsequently test those with several different inferential statistical tests (i.e., from correlation, through t-test, and logistic regression, to linear mixed models), as well as some simple psychometric ones (e.g., assess the internal consistency of groups of items). Additionally, students may explore the impact of different response formats on responses they observe (e.g., contrast binary "Yes", "No" items with items that have 11 response categories).

6.4.3 Central strengths and weaknesses

As our section on secondary analyses highlights, a core strength of our dataset is the sheer breadth of constructs covered in it. An additional strength of our dataset is that, although it can be used to investigate the relationship between newspaper consumption and held political opinions, only a minority of participants (i.e., only 11 of the 77 participants that ventured a guess) considered this relationship to be the one we likely wished to investigate with this study. This makes it less likely that demand characteristics (e.g., influences on participants' responses from them guessing our hypothesis; Ruhl, 2023) meaningfully influenced many participants' responses.

A central weakness of our dataset is that our sample was atypically (cf. Statistics Netherlands, 2024) highly educated for the Netherlands. This might partially stem from us restricting participation to people who reported having read a newspaper article the week before (which we did to ensure that each participant can fill out the first section of our questionnaire), as newspaper consumption is more common among people of higher education and social class (Elvestad & Blekesaune, 2008; Fortunati et al., 2014).Consequently, users of our dataset should be careful, when they generalize eventual statistical findings to the whole Dutch population.

A second limitation of our dataset is that it cannot be used for statistical analyses that require large sample sizes, given our modest sample size of 105 respondents, and that ratings are sparse for many individual newspapers. Results pertaining to the psychometric qualities of questionnaire items, and to individual newspapers must be interpreted with caution.

Contribution Statement

- Conceptualization: All authors
- Data curation: DB
- Methodology: DB, OvdW, MS
- Software: DB
- Formal analysis: DB
- Investigation: DB
- Translation: OvdW, LvM, DB
- Writing – Original draft: DB
- Writing – Review and editing: DB, LvM, KS
- Project administration: DB
- Supervision: WZ, LvM, KS
- Funding acquisition: LvM, KS

Conflict of Interest

The author(s) declare no conflict of interest associated with the publication of this manuscript.

Funding statement

This data was collected as part of the project "The biased reality of online media – Using stereotypes to make media manipulation visible" (project number 406.DI.19.059) of the research programme Open Competition Digitalisation-SSH, which was financed by the Dutch Research Council (NWO) in the years 2020-2026.

Chapter 7

Discussion

In the chapters of my dissertation, I present projects of my doctoral years that involved the evaluation of (different families of) measurement tools. In this final chapter of the dissertation, I will shortly discuss the conclusions I draw in regards to the research questions I posed in the Introduction – as well as open questions, caveats and other comments that provide additional context to these conclusions.

At the time of writing, it has been several years since we have written the articles presented in Chapters 2 and 3 (i.e., the most recently written sections of these are 2 years old, the oldest ones are 4). I will in this discussion chapter hence also highlight some of the ways in which my thinking has evolved in the subsequent years.[1]

Evaluation of my projects, in light of the research questions

> *Research Question 1:* What can NLP learn from psychometrics about ensuring that (bias) measurement tools capture something meaningful?

In the article that underlies Chapter 2, we made two arguments, both of which are inspired by mainstream psychometric theory on measurement and validity (i.e., construct validity; see Newton & Shaw, 2013 for an overview of how psychological validity theory evolved over time): Firstly, we argued that it is important to distinguish one's construct from one's operationalization (i.e., what

[1] I will not comment on my other chapters in a similar manner: While our empirical work, too, is several years old (e.g., for Chapter 4, I preregistered and collected data $2\frac{1}{2}$ years ago), writeup for these projects was recent.

one wants to measure from the process and tools one used to obtain scores). Secondly, we argued, that validation – the process of gathering evidence that one's interpretation of observed differences in test scores (e.g., as reflecting differences in one's intended construct) is warranted (Kane, 2013) – is an important step in the process of developing measurement tools (see also e.g., Flake & Fried, 2020).

This process of validation in essence involves exploring a lot of conditional predictions of the form "If differences in scores from my measurement instrument were reflective of differences in bias in this context, I would expect to see that ...". As guidance (and food for thought), we consequently discuss some examples of how validity and reliability of NLP bias measurement tools have been and could be assessed.

Retrospective thoughts I still stand by these two main arguments of our article – as well as by the general overarching one that NLP bias measurement can learn from psychometrics, because NLP shares some of the challenges that psychological measurement faces (e.g., the "black box" nature of its targets of measurement). This article was inspired by early influential NLP bias research, like Bolukbasi et al. (2016), which took the form of "define a metric, obtain a score, intervene on the language model, see a change in score" – with limited testing of the extent to which the newly-introduced metric was actually meaningful. Such confounding of operationalization and construct was what we tried to counteract: A number changing is only meaningful to the extent to which that reflects changes in the desired construct.

Like the older bias measurement tools that inspired us to write this article (Cao et al., 2022; Delobelle et al., 2022), also newer ones still seem to capture substantially different properties of language models: When they are used to score the same language models, their scores still are weakly correlated to uncorrelated (Asgari et al., 2026). Evidently there is still a need for conceptual work on clarifying what constructs researchers ought to measure (see also e.g., Blodgett et al., 2020) and on how they should go about doing so. I still believe that the basic psychometric vocabulary we introduced to NLPers can aid such discussions and am encouraged to see that, by now, such vocabulary has become slightly more mainstream in NLP (see e.g., Huang et al., 2024; Nguyen & Ploeger, 2025; Zaghouani, 2026).

A part of the article that, in my view, aged less gracefully is that we focused a lot on bias as an intrinsic property of language models. What we only mentioned in a footnote – that validity concerns the interpretation of scores *in a particular context and application* – deserved a bigger focus and longer discussion. I now believe that prediction of LLM behaviors in new applications is one of the largest challenges in measuring the propensity of language models to provide undesirable outputs (i.e., in measuring bias): As it is impossible to anticipate what tasks LLMs will be prompted to perform "in the wild", it will be important to ensure

that scores from bias measurement tools are predictive of the outputs LLMs will provide for a wide range of tasks – somehow also for applications of LLMs that developers of the bias measurement tool did not validate their measure for.

When we wrote the article underlying Chapter 2, LLMs and prompting were new and we did not anticipate this problem. Especially my thinking at the time was instead influenced by the earlier generations of bias measurement tools, which tried to capture internal properties of the language model (e.g., WEAT was applied to static embeddings; Caliskan et al., 2017). I took inspiration from research on personality and intelligence tests and drew parallels between humans and language models that I now see as potentially misleading: I assumed that there were likely stable characteristics of language models that – akin to personality traits – would be predictive of language models' responses in a wide variety of contexts.

This assumption may be true to some extent, as language models provide relatively consistent responses to personality questionnaires (i.e., at least across the prompt variants tried by Huang et al., 2024) and have characteristic linguistic styles (i.e., which LLM generated a text can be detected across prompts; Bitton et al., 2025). However, akin to how LLMs can show performance profiles on cognitive tasks that are atypical for humans (e.g., can perform well on tasks humans consider difficult but poorly on tasks that, to humans, are trivial), it is possible that language models, more so than humans, provide inconsistently biased responses across tasks (e.g., that a language model is largely neutral across most tasks, but provides highly sexist responses in a particular subset of tasks; or that the bias of a language model's responses for a task strongly varies for two prompts that to humans would be identical in meaning). I believe it is hence important to empirically assess the extent to which (biased) model behavior is consistent across tasks and prompts within tasks.

Additionally, given the wide potential range of tasks LLMs are applied to, I now believe that it is important to investigate which tasks a lot of people use LLMs for, and which especially consequential tasks (e.g., pertaining to hiring decisions, voting, medical decision-making, i.a.) at least some people use LLMs for. LLMs can inflict a lot of harm cumulatively with biased outputs on commonly performed tasks and a lot of harm with biased outputs to infrequent but consequential tasks. That is why I believe it is especially important to develop bias measurement tools for such tasks.

> *Research Question 2:* What can NLP learn from psychometrics about the composition of tests and about the quality of individual items?

Item response theory In the article underlying Chapter 3, we argue that statistical models from the psychometric framework of item response theory (IRT) could be applied to assess the composition of NLP bias measurement tools (i.e.,

that Test Information Curves could be used to evaluate how different levels of one's bias trait are covered by one's particular set of test items). These statistical models can also be used to evaluate the attributes of individual NLP bias test items – provided that the implemented IRT model's assumptions hold: that performance on the item is well-predicted by the assumed number of testtaker attributes (i.e., in the models we discussed: by one "model biasedness" trait), that there is enough data to adequately fit these statistical models, and that the relationship between the level of a trait a testtaker possesses and the responses it provides is monotonic in a manner similar to what the particular implemented IRT model assumes (e.g., that more biased LLMs are more likely to provide biased responses to all items; that language models have an a priori 50% chance of providing biased responses for binary items, i.a.).

Retrospective thoughts With retrospective knowledge on how important prompting would turn out to be, I would have focused the article less on benchmark datasets specifically and instead also talked about other NLP bias applications that IRT could be useful for. A particular miss here is our sparse discussion of *differential item functioning* (see Figure 3.7) of test items between different groups of testtakers. I believe that NLP bias measurement tools could be developed inspired by this notion – for example by comparing a LLM's performance on test items with persona prompting that is identical except for indicators of gender (e.g., "You are a [adjective] physician. Please answer the following questions consistent with your role: [...]", with "[adjective]" being "female" vs "male", respectively).

In regards to what we did write about, it is worth acknowledging that this work was purely theoretical with no empirical demonstration that IRT models can be readily applied to NLP bias measurement. The jury is still out on whether they can be. As we described in the article, a potential hurdle is the large number of testtakers that are required to fit IRT models. Bayesian fitting with informative prior distributions could decrease the number of required testtakers, but, as I described in regards to the previous research question, it is an open empirical question to what extent assumptions that commonly hold for human testtakers (e.g., that many of their traits are approximately normally distributed; that relationships between levels of these traits and performance on tests assessing these traits are monotonic, i.a.) also hold when the analyzed "testtakers" are LLMs. If these assumptions are violated for LLMs, that could be problematic for fitting IRT models. For one, this could make it necessary to test many LLMs, as it could be impossible to find prior distributions that describe the attributes of LLMs well (ruling out Bayesian fitting). For another, in the worst case scenario, where LLMs' responses are not well-described by monotonic relationships between traits and performance, or by a small number of traits (e.g., if there is no overall "propensity to provide undesirable responses" that LLMs differ on),

IRT models are largely inapplicable to NLP bias measurement.

That being said, even if actual IRT models do not find application, I believe that the IRT framework is a useful conceptual lens to be aware of for NLP bias researchers: that more items in a test does not equal a better test, that tests should be assessed by their composition (e.g., in regards to the different bias severities they can be used to detect), and that LLM testtakers' responses to items are best thought of as probabilistic (conditional on their LLM's level of bias) are all points that I believe were worth making. Item response theory provides a good framework to consider these points and offers some heuristics about what good tests should look like (e.g., in regards to their composition).

Silicon sampling In the appendix to Chapter 6, I discuss an alternative psychometric approach to assessing the attributes of individual test items and of assessing the composition of tests. This involves a two-step process: First one specifies one's assumptions about the relationship between unobservable latent variables (which one generally assumes to be the constructs one intended to measure with one's items) and observed test items in a Structural Equation Model. Subsequently one tests whether observed correlational patterns between testtakers' responses are consistent with these assumptions. As we explored whether to practically apply silicon sampling approaches, we were especially interested in the extent to which measurement invariance holds, that is: the extent to which the relationship between test items and latent variables was identical between human respondents and their LLM imitators.

The pilot study in Chapter 6 does not allow me to answer this question, yet. We did not only want to investigate this question of measurement invariance, but also the extent to which silicon samples can accurately depict the relationship between people's socio-demographic variables and their political beliefs. As our data was not promising for assessing that second question (i.e., we found virtually no relationship between our respondents' socio-demographics and political beliefs) and since that question was more important for our overall project about people's media consumption and political beliefs, we did not proceed past this piloting stage. Consequently we did not collect enough data to assess the question on measurement invariance with structural equation models (which require data from hundreds of participants).

I still believe that this approach to comparing humans with their LLM imitators is promising. That being said, measurement invariance between people and their silicon samples could be better evaluated with a study solely focused on that goal (cf. our dual purpose setup). Then one could obtain silicon sample predictions for already existing and sufficiently large datasets, like the one of the European Social Survey (European Social Survey European Research Infrastructure, 2024) which includes data of $\approx$ 1500 humans respondents, each, for many of its countries. Especially such multinational datasets are promising, as one can

use them to test LLMs' capacity to imitate human respondents outside of the US-American context for which most state-of-the-art LLMs have been designed (which is also the main political context in which silicon sampling has previously been applied; von der Heyde et al., 2026).

Additionally useful would be to assess the extent to which measurement invariance holds for other psychological constructs that demographics are thought to be predictive of, like personality traits (e.g., neuroticism is generally higher in women than men; Costa et al., 2001; and generally decreases with age; Bleidorn et al., 2022) and religiosity (i.e., religiosity generally increases with age; Bleidorn et al., 2023; and by several different metrics – though relationships differ by religion – women are generally more religious than men; Pew Research Center, 2016). Thereby one can evaluate whether imitating people's response patterns is a general capacity of state-of-the-art LLMs or one that is limited to particular applications or cultural contexts.

> *Research Question 3:* If we ensure that participants are unaware of what is being tested, can we still find evidence of sequential gender stereotype priming?

Conditional on a few important caveats, my empirical results in Chapter 5 point towards "no": We found no evidence for stereotype priming, independently of whether we analyzed participants' accuracy or reaction times (and across different ways of excluding individual trials and participants). Not only the ultimate results of our iterative series of model comparisons but also their sequences suggest that no gender stereotype priming took place: (with one exception, each) every predictor related to the gender stereotypicality of words was excluded before any other predictor, indicating that these gender associations performed the worst at explaining participants' performance.

General caveats about the nature of probability and one-off results aside (e.g., we might have just gotten very unlucky and sampled highly atypical participants and/or stimuli resulting in misleading results), there are at least two sets of caveats to qualify that conclusion: The first, which we discuss in detail in the Discussion section of the chapter, concerns my particular implementation of a gender priming paradigm (e.g., that we ran an online rather than a laboratory study, and that we let participants perform novel tasks which might have made priming difficult to detect; see the Discussion section). A second caveat (not discussed in the chapter) concerns my statistical power to detect priming effects. While, we had a large number of participants and trials compared to previous studies that are said to provide evidence for gender stereotype priming (e.g., Banaji and Hardin 1996's effect is based on 60 participants who each performed 80 relevant trials; and Karylowski et al. 2001 had 31 participants, times 288 trials), in absolute numbers 100 participants and 160 trials per participant still does

not afford a lot of statistical power to detect what is probably small differences in reaction time and/or accuracy between stereotype-congruent and stereotype-incongruent trials. If gender priming effects exist but are tiny, the amount of data we collected was likely insufficient to detect them.

That being said, if effects are indeed this small, sequential stereotype priming is also not a particularly useful class of measurement instruments for our intended project: What we tried and failed to detect in Chapter 5 is the existence of any difference in the attributes people ascribe to women versus men. Even more statistical power would be required to make meaningful claims about individual differences (e.g., that person 1, based on these measures, has stronger gender associations than person 2), or – which is how we would have applied priming paradigms to study the effects of newspaper consumption – about changes over time in people's gender associations. While I did not test this assumption here, I suspect there would have been statistically discernible differences in the agency- and communion-related traits that people ascribe to women versus men (given the large differences that are often found for such traits, see e.g., Eagly et al., 2020; Haines et al., 2016) – had we collected questionnaire data, instead of priming data, for the same number of participants.

> *Research Question 4:* How much are people's reported views on the 'genderedness' of occupations influenced by how the question is posed?

This question I wondered before we collected the dataset that I report on in Chapter 4. I expected that the task of "gendering" different occupations is controversial and subject to demand characteristics (e.g., people responding in "politically correct" ways that are inconsistent with their genuine beliefs). Consequently, I assumed that the framing of the question (i.e., whether we asked participants about their opinions or about what they perceived to be their country's consensus opinion) and the number of response options (i.e., even or odd) would have a large effect on what responses participants provided. In a sense it did: For the groups of people that had an odd number of response categories, the neutral (ungendered) middle option was the default response that received a plurality of responses.

In regards to the overall trends however, results were much more similar than I expected: Among all tested populations, the questionnaire versions had close to identical results for which occupations were overall rated as masculine versus feminine and for how these occupations ordered on the masculine-to-feminine scale.[2] For our intended use of this dataset, as a validation tool for NLP bias

[2]If one is interested in which occupations are strongly gender stereotyped, rather than in which direction they are stereotyped (i.e., whether they are stereotypically feminine or masculine), results were additionally robust, independently of whether one defined occupations as "strongly stereotyped" by how far their mean ratings diverge from a neutral midpoint, or by

measurement instruments, this is good news: All my questionnaire versions seem to converge on the same overall sentiments of US and UK adults about different occupations. Had conclusions of these different versions substantially diverged, it would be difficult to advise the creators of NLP bias measurement tools on how they should use the data (e.g., data from which of the different questionnaire versions they should use, or whether to aggregate some or all of the versions' data).

A few (potential) weaknesses of our dataset are worth noting. The first concerns the implementation of the Shiny application for curating and downloading the dataset: The application automatically aggregates (not-filtered-out) data through means and standard deviations. Mean aggregates compress information and are potentially misleading, for example in cases of bimodality (e.g., if there is a subset of participants that rate a particular occupation as highly masculine and another subset that rates the occupation as highly feminine, a mean aggregate will hide this tension and instead show a value between these two extremes). Since NLP bias researchers seem to want singular numbers that indicate occupations' gender stereotypicality (as evidence by them often using demographic statistics like "percentage of women in a given occupation") and as such default options make the dataset easier to use, I believe that a mean default is defensible.That being said, researchers that care especially about a particular subset of occupations rather than about overall trends across occupations are advised to download the raw data and explore the response distributions for these occupations, in detail.

The second set of weaknesses concerns the nature of our data. For one, because our data is in English and treats gender as binary – when NLP gender bias measurement tools have been criticized as too binary and too focused on English (Stanczak & Augenstein, 2021) – we indirectly encourage the creation of more such bias measurement tools, by providing validation data specifically for them. We chose to treat gender as binary, as it allowed us to replicate existing social psychological studies which treated gender as binary (i.e., Beggs & Doolittle, 1993; Munoz Sastre et al., 2000; Shinar, 1975; White et al., 1989). An additional reason to provide binary English-language gender data is that most NLP gender bias measurement tools share exactly these (potentially flawed) attributes: In the absence of our dataset, many measurement tool creators would resort to occupational gender demographics, so we addressed an urgent "gap in the market".

A final weakness of our dataset is that for some more obscure occupations, data is sparse from participants that know about them: For example, only 24 UK participants were at least moderately familiar with the occupation "arbiter". While we warn in the chapter against putting much stock into mean ratings for occupations with such sparse data, there is a risk that such occupations become standard fixtures of NLP gender bias measurement tools, since validation data

how few of their ratings are "neutral" / non-gendered (see Section 4.4.4).

for them – our data – exists.

Concluding thoughts

As scientists we are among a privileged few people who can afford the time to wrestle with some of the big questions in life. For many of these questions, (testing and) observation is necessary to make progress. These "empirical" questions require some form of measurement, a way of compressing the near-infinite complexity of the outside world into a finite and hence tractable set of data, often in the form of numbers.

In this respect, measurement resembles other attempts of capturing reality: Like a painter in a wild forest, we cannot capture nature in its full glory, but must decide what to "paint", and how to approach "painting". To be able to measure, we make choices about which parts of reality we attempt to represent (and how we conceptualize them); about which details we abstract away because we consider them irrelevant (Aristotelian idealizations; Frigg & Hartmann, 2006); and about which details we must abstract away for practical reasons, even though we consider them relevant (Galilean idealizations; Weisberg, 2007). Akin to a painting, our representation of reality is necessarily imperfect; another scientist looking at the same wildlife scene might leave it thinking that we omitted or distorted features that were critical to understanding the natural phenomenon.

Much of my doctoral work, reflected in this dissertation, was devoted to this "art form" of measurement. In the projects of the first part of this dissertation, I tried to bridge two research traditions, Natural Language Processing and Psychometrics, as I believed the former to ask pressing new questions (i.e., about the potential societal impact of widespread LLM use), and the latter to have tools for approaching these questions. In the second part of my dissertation, I evaluated the assumptions behind a popular approach to measurement (sequential stereotype priming), and attempted to evaluate those of a novel one that is swiftly becoming more popular (silicon sampling). In essence, I wondered whether what appeared on other painters' canvases reflected the original wildlife scene, or was (partly) an artifact of their painting technique.

As painters have personal styles, so do scientists. We differ in which imperfections in our – or other people's – paintings bother us enough to intervene. I, too, am inconsistent in what I scrutinize: Some of my prior assumptions – like that scale order would shift the distribution of respondents' responses, when measuring the "genderedness" of occupations (an assumption that turned out incorrect; see Chapter 4) – I actively tested; others sounded sufficiently plausible to me to provisionally accept. For instance, I designed the questionnaire in Chapter 6 with the assumption that awareness of a research question can impact participants' responses – for example, that participants might provide less politically radical responses to protect their favorite newspapers, if they knew that we were

trying to test the link between newspaper content and political opinions. Considerable effort (i.e., about two thirds of the length of the questionnaire) was hence spent on concealing our research aims. Yet, if questionnaires are less susceptible to socially desirable responding than commonly thought, as has recently been argued by Corneille and Gawronski (2024), the cost of including such distractor items (e.g., they could confuse participants, or drain their attention) might not have been worth it. If this is true, we were additionally unwise to exclude data from participants who correctly guessed our research aims. Another important assumption I hold but have not explored much, is that, when applying a psychometric lens to NLP bias measurement, individual language models are the appropriate analogue to human testtakers. Many of my suggestions (e.g., about applying statistical models from item response theory) are conditional on this mapping being appropriate.

In addition to such "known unknowns" – assumptions that we scientists recognize as worth exploring in principle but chose not to – there are also assumptions whose uncertainty we fail to recognize (e.g., because we unduly consider them settled science), and "unknown unknowns": assumptions that we hold without realizing. With many years of hindsight, as when one nowadays reads Goddard (1917), it is easy to find examples of such hidden and potentially pernicious assumptions. Shaped by the norms of our time (as was Goddard in his time; Gelb, 1986) and our personal experiences, we find it harder to detect these types of assumptions in contemporary work – and hardest still to detect them in our own work. If there is anything, dear reader, that you take away from however many words of this dissertation you have read, I hope it to be the following: It is worth occasionally revisiting, explicating, scrutinizing, and inviting scrutiny of your assumptions about your measurement tools and measurement processes.

Appendix A

Appendices for Chapter 4

A.1 [S] Changes from our preregistration

Additional excluded participants (Section 4.2.2)

- We excluded participants who did not fully complete the study. We did not preregister this, because we were unaware that Qualtrics saves partial data. This led to the exclusion of 4 participants (2 from the US-convenience sample, 1 from the US-representative and 1 from the UK-representative sample) who would have passed the preregistered exclusion criteria.

Test we forewent (Section 4.4.1)

- We attempted to fit the full factorial ANOVA that we preregistered to test which factors were predictive of rating differences, but the model a) took too long to compile to make it feasible to fit several simpler versions of the model for computing inclusion Bayes Factors and b) ultimately did not converge.

  Section 4.4.1's parallel pipelines of exploratory analyses were inspired by our intended ANOVA, with three major differences: Firstly, we no longer tested for the presence of differences between countries but assumed their existence and just estimated their size (by including random occupation-based "sample" slopes). Secondly, we only indirectly test whether the 6 vs 7-point scales differ, by running parallel analysis pipelines. Thirdly, we only assessed main effects. All these changes decreased our model space and model complexity, making the computation of inclusion Bayes Factors more feasible.

Forgotten occupation (Section 4.4.2)

- For the comparison between our ratings and the ratings of Shinar (1975) and White et al., 1989, we preregistered that we would compare their ratings for "journalist" and "reporter" – which White et al. saw as equivalent terms – with a mean of our stereotypicality ratings for these two occupations. We had planned on collecting separate ratings for these two terms, because in our view they describe slightly different occupations (i.e., journalist is a wider category than reporter). However, we accidentally did not solicit ratings for "reporter", only for "journalist". Consequently, we compared their ratings to our rating for "journalist", only.

Typo (Section 4.4.3)

- In our preregistration of the Bayesian correlation analysis, we refer to a paper with doi 10.3758/s13428-015-0593-0 . That is the wrong doi (its article has nothing to do with statistics) and was supposed to refer to Wagenmakers et al. (2018), a paper introducing Bayesian correlation tests in JASP (JASP Team, 2025). We hence used that paper's approach – consistent with what we intended to preregister.

A.2 Appendix: [S] Differences between Shinar's study and our replication

Questionnaire versions / instructions In Shinar (1975)'s original study, she compared ratings that participants gave on identical response categories, but in response to three different instructions: Besides the rating-scale-explaining instructions (Shinar's Group 1) that we replicated here (see Figure 4.3 I.A), other participants were asked to rank occupations by what they perceived to be the proportion of men and women employed in them (Group 2), while yet others (Group 3) responded based on "the personality characteristics of people employed in them" (e.g., to rate, as her instruction read, "occupations typified by predominantly feminine traits" as feminine). Earlier partial replications with US samples made different choices: White et al., 1989 replicated Shinar's second and third instructions, with (at least, as far as they mention) only the "masculine" to "feminine" scale direction, while Beggs and Doolittle (1993) used neither of the three and instead provided vague "use your own judgment" instructions.

While we do believe it would be interesting to see whether Shinar (1975) and White et al., 1989's finding – that people across these instructions provided highly similar ratings – replicates, we chose to only include her first set of instructions. Firstly, because their findings imply that the different instructions could be somewhat redundant which made it hard for us to justify including 4 additional questionnaire versions (i.e., 2 additional instructions x 2 scale directions).

Secondly because we believed that Shinar's first instruction more directly tabs the "gender stereotypicality" construct we aimed to measure here: While participants' views about the gender proportion in different occupations and about personality traits typical for different occupations are likely skewed by the extent to which they see occupations as gendered, her first set of instructions leaves no need for such indirect inferences: There she just directly asked her participants about whether they see occupations as gendered.

Criterion follow-up question We made three small changes to Shinar (1975)'s follow-up question about the criteria participants used to judge the gender stereotypicality of occupations. Firstly, we changed the fourth answer option from "Other" to "On the basis of something else", as we considered the latter less ambiguous. Secondly, we ask participants which chose that fourth answer to specify their "other" criterion. Finally, whenever participants chose more than one answer option on that question, Shinar (1975) would ask participant which of their criteria were primary and secondary. Instead we ask participants to estimate how much each of the different chosen criteria weighted in their judgment, as this indirectly provides Shinar (1975)'s "primary" vs "secondary" rankings.

We believe these changes provide us with additional information – for example, about much more important a "primary" criterion was compared to the "secondary" criterion – at virtually no cost[1]: As the question about decision criteria succeeds the stereotypicality ratings, our stereotypicality ratings are directly comparable with those of Shinar (1975), despite the follow-up questions slightly differing.

A.3 Appendix: [N] [S] Ethnicities of the samples

Prolific provides access to demographic data for all 954 participants that attempted participating. This includes data for participants which opened the study but quit (e.g., at the very beginning because they do not meet the inclusion criteria; or during the study because they got bored). Instead we can also match our participants' experimental data to their ethnicity through Prolific's internal ID codes – though we have access to these for only about half of our participants. That means we have two flawed estimates for the ethnic makeup of our final samples, here: One for all 954 participants that attempted our study –

[1] If participants choose the same percentage values for some of the criteria (e.g., 40-30-30, if they previously indicated that 3 criteria apply), we arguably lose a bit of information compared to Shinar (1975), as her participants had to tie-break and pick a primary and secondary criterion. Indeed 194 of our 619 multi-criteria participants gave the same percentage scores to two or more criteria. That being said, knowing when criteria weighed about equally in a decision is also information which Shinar's approach would miss.

that is: 175 more than what our rating tables and analyses are based on – and one for 404 participants that we could match with our final participants.

Table A.1 contains counts for the former, Table A.2 for the latter. Counts are shown by self-reported ethnicity (top row) and sex (second row). Entries with "NA" represent missing data (e.g., because participants refused to answer that question).

When contrasting the US convenience sample with the US sample for which we paid an additional fee to make it more representative, there seem to be few differences in ethnic diversity. The representative sample however has more female than male participants while the reverse is true for the convenience sample. Additionally, participants in the representative sample, at least post exclusions, tended to be older (see Section 4.2.2).

Table A.1: Ethnicity and sex of all 954 people who attempted our study

	Asian		Black		White			Mixed		Other		NA	
	F	M	F	M	F	M	NA	F	M	F	M	M	NA
UK	10	8	5	5	131	123	0	5	5	5	5	0	14
US_{rep}	10	9	21	18	114	112	0	6	5	5	5	0	16
US_{conv}	8	12	14	23	105	118	2	5	6	2	5	3	14

Table A.2: Ethnicity and sex for 404 of our final 779 participants

	Asian		Black		White			Mixed		Other		NA
	F	M	F	M	F	M	NA	F	M	F	M	NA
UK	5	3	1	4	55	67	0	4	1	2	3	0
US_{rep}	5	3	13	7	49	43	0	3	3	2	3	1
US_{conv}	0	3	5	7	53	46	2	1	6	1	3	0

A.4 Appendix: Details on the Bayesian analyses

Unless otherwise specified, Bayesian models reported in this paper were fit with brms (Bürkner, 2017), and Bayes Factors for model comparisons were based on bridge sampling (Grün & Bürkner, 2020). Below, we describe prior distributions (and, if applicable, convergence diagnostics) for all analyses, in more detail.

[S] Differences between questionnaire types in the "physical capabilities" criterion (Section 4.3.1) Seventy-nine percent of participants completing the "Social Desirability" questionnaire versions reported that people in their country make judgments about the femininity/masculinity of occupations based on the "physical capabilities required for the occupation", compared to only 69.6% of participants (who completed the Shinar versions of the questionnaire) who reported using this criterion themselves.

We performed (exploratory) logistic regression analyses to test whether this difference is worth a mention. A logistic regression that included questionnaire type as a slope explained the participants' rates of choosing the "physical capabilities" criterion 5.8 times better than a model that only included the intercept.[2] This suggests that the observed difference in endorsement rates is meaningful – though it should be taken with a grain of salt: While the pattern is consistent across samples (UK: 75.2 vs 65.9%, US-convenience: 81.9 vs 69.3%, US-representative: 80.3 vs 74.1%), we had 4 criteria, 3 samples, and 2 questionnaire types to inspect. There were hence several factor combinations which could produce such a "noteworthy" pattern by chance.

For this analysis, we used normal distributions with means 0 and standard deviations 5 as prior distributions for the intercepts of both models. For the "questionnaire type" slope we used a normal with mean 0 and standard deviation 2.5 (this is based on Gelman et al. 2008's weakly informative default Cauchy prior). Across 4 chains, we each drew 2000 warmup and 2000 post-warmup samples (no thinning). Convergence was sufficient, all $\hat{R}s = 1$, all bulk and tail effective sample sizes ≥ 2768.

[P] Analysis of near-exclusive 3 and 4 responders (Section 4.3.1) Five participants that responded to 6-point scales provided "4 - somewhat feminine" and "3 - somewhat masculine" responses on almost all (i.e., $\geq$ 95%) of their trials. We were curious whether these responses still contained signal – that is: were consistent with others' stereotypicality ratings – or whether they were mostly

[2]For readers that prefer to draw inferences based on posteriors instead of BFs: The posterior mean of the (questionnaire type) binary slope parameter is -0.5 in log odds, with an equal-tailed credible interval of [-0.83, -0.17].

noise (e.g., since no neutral response option exists, these participants randomly alternate between the mid-most options).

We consequently, compared two Bayesian hierarchical logistic regression models, both of which tried to predict whether a participant would respond with a 3 or a 4 on a trial: The first model had a fixed effects intercept and random intercepts for participant and rated occupation. The second had an identical structure (and priors) but also included a fixed effects slope for ratings of occupations from other participants. Specifically, for these ratings, we calculated means based on the data of all (non-excluded) participants that completed the 6-point scale – except for the 5 participants that were assessed here – on all occupations besides the catch ones "nun" and "pope". Subsequently, to simplify the choice of priors, we standardized ratings and used these standardized ratings as predictors on each trial, depending on the trial's to-be-assessed occupation. To enable the binary contrast between response option 3 vs 4, we excluded trials from our five participants on which they chose other response options. This led to the exclusion of 33 out of 1040 (i.e., 208 occupations times 5 participants) responses.

The posterior mean of the slope parameter indicates that an increase of 1 standard deviation in other people's mean rating comes with an increase of 2.43 in log odds (equal-tailed credible interval: [2.10, 2.80]) that participants will choose response option 4 rather than 3. That is a substantial effect, as becomes evident if one, for a moment, idealizes away posterior uncertainty and random effects, and contrasts predictions, based solely on the posterior means of the fixed effects intercept (i.e., −1.26) and slope, for the occupations rated most and least masculine: The predicted probability that 4 is picked is $p(-1.75) = \frac{1}{1+e^{-(2.43(-1.75)-1.26)}} = 0.4\%$ for miner and $p(2.59) = \frac{1}{1+e^{-(2.43(2.59)-1.26)}} = 99.4\%$ for beautician. Unsurprisingly, model comparison overwhelmingly favored the model that included the slope over the model that did not, $log(BF_{ModelComparison}) = 124.66$.[3]

For this analysis, we used normal distributions with means 0 and standard deviations 1 as prior distributions for the fixed effects of both models (i.e., fixed intercept for both models, fixed slope for "standardized mean occupation rating" for one of the models). For the random intercepts, both models had t-distribution priors with 3 degrees of freedom, means of 0 and standard deviations of 1.

We drew 1000 warmup and 3000 post-warmup samples across 4 chains. Convergence was sufficient, all $\hat{R}s = 1$, all bulk and tail effective sample sizes ≥ 1913.

[N] [P] [S] **Determining which factors influenced stereotypicality ratings (Section 4.4.1)** To determine whether or not four different factors, participants' gender, age, residency (UK vs US) and scale direction (feminine first vs

[3] As a sanity check to ensure this result did not stem from misspecification (e.g., of the slope's prior), we also compared equivalent Frequentist hierarchical logistic models. Again we founds overwhelming evidence in favor of the model with the slope, $\chi_{ModelComparison}(1) = 270.01, \triangle BIC = 270.02$.

masculine first) influenced stereotypicality ratings, we fit 16 Bayesian hierarchical models per scale type (i.e., 6-point scales, and 7-point scales). These models differed in their fixed effect structure – each model include one of the 16 possible combinations of the four factors as fixed main effects – but had an identical random effects structure: a random intercept for participants (which accounts for baseline differences in ratings between participants) and for occupation (which accounts for baseline differences in ratings between occupations) and a random slope for residency based on occupation (which accounts for differences between UK and US residents in how they rate different occupations).

After fitting the models, we used bridge sampling to receive estimates of the models' marginal likelihoods. For each factor, we subsequently computed inclusion Bayes factors by comparing the summed marginal likelihoods of models including the factor to those excluding it (weighted by prior model probabilities, which we assumed to be uniform).

We used identical prior distributions for all models within and across the two model series. As the models predicted a standardized dependent variable, we used priors that were weakly informative on that scale: All parameters of all models either had a Cauchy or (for bounded parameters like SDs) half-Cauchy prior with $mean = 0$ and scale $r = 0.354$ – priors that treat the absence of effects as most plausible, small to medium effects (e.g., slopes between 0.1 to 0.5 – which on the standardized scale correspond to small to medium-sized effects) as plausible and more extreme values as possible but unlikely. For the random effects correlation matrices, we used LKJ(1) priors, uniform priors which indicate that we are a priori agnostic about what kind of random effects correlations we could observe.

We drew 2000 warmup and 8000 post-warmup samples across 4 chains for each of our models. Convergence was fine: The maximum $\hat{R}$ was 1.01, and the bulk and tail effective sample sizes were $\geq$ 466 for the fixed effects intercepts (the parameter with the consistently lowest bulk ESS), and $\geq$ 1148 for all other parameters.

[S] **Comparisons with previous studies (Section 4.4.2)** The comparison between our ratings and the ratings from Shinar (1975) and White et al., (1989) involved 3 series of analyses.

Firstly, we tested whether it was reasonable for us to pool ratings across our two scale directions (i.e., across our scales that started versus ended with the feminine half) for this analysis series. To that end, we compared extremity scores (i.e., the number of scale points an occupation mean was away from the neutral 4 midpoint) between our "feminine first" and "masculine first" samples with a Bayesian t-test. We performed this analysis with the BayesFactor package (Morey & Rouder, 2023) which analytically computes a Bayes Factor which evaluates the evidence for a difference versus no difference in extremity score group means. As preregistered, we used weakly informative JASP default priors for all our analyses:

In this case, it was a Cauchy with scale $r = 0.707$. As we only performed this analysis to evaluate whether pooling is reasonable (which it was; $BF_{01} = 3.59$), we did not sample from the posterior.

Subsequently, we ran two preregistered analyses of variance, both times predicting occupations' extremity scores by study membership (i.e., ours vs Shinar vs White et al.) – though differing in which occupations were compared. We once more performed these analyses with the BayesFactor package with JASP default priors (i.e., their default "fixed effects" prior for ANOVAs): Cauchys with scale $r = 0.5$. We once more primarily cared about whether the analytical Bayes Factor would support or oppose the existence of group differences and hence did not sample from the posterior.

Finally, as exploratory analyses, we ran two Bayesian hierarchical models with brms. These predicted extremity scores based on a random intercept for occupation, a fixed effects intercept, and – for one but not the other model – a fixed effects slope for study membership. The models had identical prior distributions centered at 0: Student-t distributions with 3 degrees of freedom and a scale of 1 for the random effects and residual standard deviations, and Gaussian priors with a standard deviation of 0.5 for the fixed effect interval and study membership fixed effects predictor. The latter embodies a belief that group differences of up to 1 extremity score are plausible, while bigger differences are unlikely. Given that extremity scores are bounded to a range of [0, 3] (e.g., the minimum [maximum] possible occupation mean of 1 [7] would be 3 points away from the midpoint), we considered this range reasonable.

We drew 2000 warmup and 2000 post-warmup samples across 3 chains for our two hierarchical models. Convergence was sufficient: The maximum $\hat{R}$ was 1.01, and the bulk and tail effective sample sizes were ≥ 631 for the fixed effects intercepts (the parameter with the lowest ESS in both models), and ≥ 1019 for all other parameters.

[N] [S] **Correlation between Stereotypicality Ratings and percentage of women working in the occupation (Section 4.4.3)** To evaluate the extent to which occupational gender participation rates can be used as proxies for how stereotypically feminine or masculine occupations are viewed, we performed a Bayesian Pearson correlation test on ranks of our (standardized and pooled) stereotypicality ratings and ranks of the percentage of women working in that occupation. Since Spearman's rho is identical to the Pearson correlation of ranked variables – provided that, as we did here, ties in rank are resolved by assigning the average rank to all tied observations – we in essence performed a Bayesian Spearman's rho test. Additionally, as exploratory analyses, we performed an equivalent Pearson correlation test on non-ranked version of these variables. For all 2 (UK and US) x 2 ("Spearman" vs Pearson) tests, we used the BayesFactor package (Morey & Rouder, 2023) and set wide beta priors (i.e., with scale parameters

of $r = 1$; see Ly et al., 2018), under which both low and high correlations are plausible.

All credible intervals reported in the main text are each based on 10000 posterior draws. Repeated sampling led to negligible numerical differences (i.e., at the fourth decimal place of the credible interval bounds), suggesting low Monte Carlo error. Unsurprisingly, given the high parameter estimates, the correlation tests all overwhelmingly supported the alternative hypotheses that correlations are non-zero; all (analytically computed) BFs $\geq 10^{20}$.

[P] [S] **Correlations involving the percentage of midpoint responses (Section 4.4.4)** To evaluate the extent to which midpoint responses reflect unfamiliarity with an occupation, we performed two Bayesian Pearson correlation test – one for UK, and one for US participants – between each occupation's percentage of midpoint responses (on the 7-point scale) and their mean Familiarity rating among 7-point scale respondents. Additionally, to evaluate the extent to which non-midpoint responses indicate strong stereotypical beliefs, we performed two Bayesian Pearson correlation tests (one per nationality) between an occupation's percentage of midpoint responses (on the 7-point scale) and the extremity of their mean response on the 6-point scale (i.e., an absolute value of how far away its mean was from the neutral mean of 3.5).

For all these 2 (UK or US) x 2 (Familiarity or Extremity) correlation tests, we used the BayesFactor package (Morey & Rouder, 2023) and set wide beta priors (i.e., with scale parameters of $r = 1$; see Ly et al., 2018), under which both low and high correlations are plausible. We also excluded data for "nun" and "pope" from all these analyses, as they are unlikely to have midpoint responses (i.e., a participants' data is excluded if they provide such responses, unless they also indicate not knowing these occupations; see Section 4.2.2). All credible intervals reported in the main text are each based on 10000 posterior draws. Repeated sampling led to small numerical differences (i.e., usually at the third, sometimes at the second decimal place of the posteriors' quantiles), suggesting Monte Carlo error of too small a magnitude to impact our conclusions.

The correlation tests provided anecdotal evidence for a relationship between familiarity and midpoint responses in our UK sample, (analytically computed) $BF_{10} = 2.05$, and against such a relationship existing in our US samples, $BF_{01} = 2.68$. They further provided overwhelming evidence for the existence of a negative correlation between an occupation's percentage of midpoint responses (on the 7-point scale) and the extremity of its average 6-point scale response, both $BF_{10} > 10^{61}$.

A.5 Appendix: [S] Table of ratings for occupations shared with one previous study

Table A.3: Means and standard deviations of gender stereotypicality ratings from our US convenience sample with the 7-point response scale, as well as either White et al., 1989's, or Shinar (1975)'s and Beggs and Doolittle (1993)'s ratings on these occupations with the same scale – depending on which of the studies collected ratings for a particular occupation. While White et al. have differing numbers (see N_{White}), each of our means is based on 127, each of Shinar's on 120, and each of Beggs and Doolittle's on 142 participants. Occupations are sorted from "masculine" to "feminine" based on our ratings. No standard deviations are listed for Beggs and Doolittle, as they did not report these.

Occupation	Bachmann	White	N_{White}	Shinar	Beggs
Miner	1.50 (0.82)			1.00 (0.00)	1.79
Pawnbroker	2.51 (1.11)			2.00 (0.73)	2.31
Magician	2.65 (1.05)			3.08 (0.81)	2.48
Mayor	2.79 (1.12)			1.83 (1.24)	2.42
Surgeon	2.88 (1.12)			2.17 (1.42)	2.60
Astronomer	2.94 (1.15)			2.92 (1.72)	2.97
Statistician	3.13 (1.16)			3.42 (0.99)	3.35
Composer	3.16 (1.02)			2.92 (0.36)	3.25
Comedian	3.16 (1.13)			3.75 (0.39)	3.35
Meteorologist	3.26 (1.07)			3.17 (1.06)	2.91
Lawyer	3.30 (1.11)	3.07 (0.96)	177		
Chef	3.31 (1.15)	2.93 (1.19)	177		
Agronomist	3.39 (1.05)			2.50 (1.54)	3.04
Banker	3.44 (1.24)			2.50 (1.00)	2.91
Broadcaster	3.68 (1.02)	3.53 (0.91)	159		
Biologist	3.72 (1.07)	3.26 (0.87)	177		
Musician	3.74 (0.80)	3.87 (0.83)	177		
Psychiatrist	4.03 (0.91)			3.67 (0.42)	3.75
Photographer	4.05 (1.15)	3.71 (0.63)	177		
Writer	4.09 (0.78)			3.92 (0.08)	4.09
Psychologist	4.28 (1.23)	3.72 (0.63)	177		
Sociologist	4.48 (1.14)	4.00 (0.84)	175		
Singer	4.51 (0.86)			4.50 (0.46)	4.19
Florist	5.72 (1.14)	5.61 (1.17)	176		
Librarian	5.74 (1.10)	5.86 (1.07)	176		
Receptionist	6.06 (1.07)			6.33 (0.79)	6.09
Secretary	6.16 (1.11)	6.23 (1.00)	176		
Manicurist	6.23 (1.16)			6.67 (0.24)	6.24
Beautician	6.40 (0.88)	6.08 (1.13)	176		

A.6 Appendix: [N] Overview of the OccupatiEN Shiny app's user interface

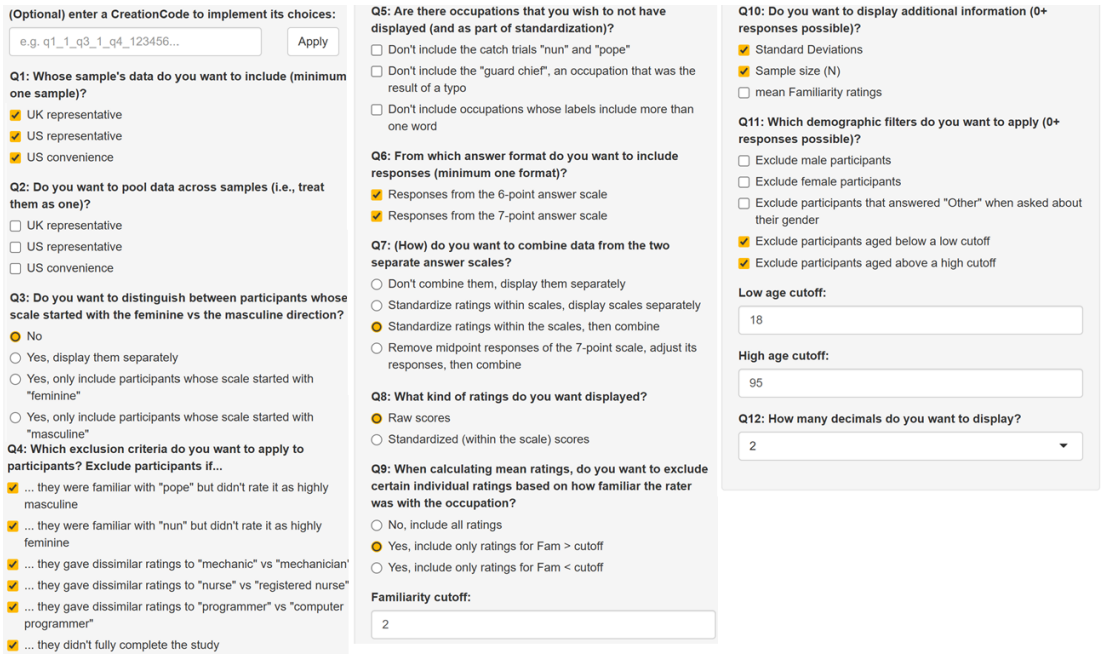


Figure A.1: Left panel of the OccupatiEN R Shiny App. Users answer these 12 questions to customize the dataset, which updates the downloadable .csv file and its preview table. As different choices lead to different ratings, the code in the column "CreationCode" saves which choices were made. Within the code, q[Integer] indicates whose question's responses are following. For example, `q1_13_q3_2_...` indicates that the user chose the first and third options for Q1, no option for Q2, and the second option for Q3. Entering a valid code in the field above Q1 reproduces a configuration of choices (e.g., `q1_3_q3_1_q4_123456_q6_2_q8_1_q9_1_q10_12_q12_2` implements the choices underlying Tables 4.2 and A.3).

A.7 Appendix: [S] Gender Stereotypicality Ratings

Table A.4: Ratings of occupations on the 7-point scale (with 1 = Masculine, 4 = Neutral, 7 = Feminine), sorted alphabetically. Mean ratings and standard deviations stem from the N-sized subset of participants which indicated being at least "moderately familiar" with the occupation. Mean familiarity ratings (Fam.), on the other hand, stem from all 243 US and 132 UK 7-point scale participants. We display ratings from the 7-point scale for historical continuity with Shinar (1975), White et al., 1989, Beggs and Doolittle (1993), and Munoz Sastre et al. (2000). An equivalent table for 6-point scale ratings can easily be created with the OccupatiEN Shiny app (see Section 4.3.2).

	US			UK		
Occupation	Mean (SD)	N	Fam.	Mean (SD)	N	Fam.
academic	3.88 (0.89)	202	3.71	3.76 (0.68)	104	3.67
accountant	3.52 (1.26)	208	3.95	3.43 (0.91)	109	3.77
actor	3.40 (1.24)	213	4.13	3.60 (0.98)	100	3.76
administrator	3.87 (1.13)	202	3.74	4.68 (0.82)	106	3.73
advisor	3.63 (0.98)	189	3.53	3.91 (0.68)	87	3.05
agronomist	3.03 (1.13)	65	1.94	3.48 (0.67)	23	1.60
analyst	3.52 (1.08)	170	3.37	3.53 (0.84)	79	2.88
animator	3.43 (1.04)	166	3.47	3.70 (0.64)	74	2.84
appraiser	3.26 (1.05)	166	3.22	3.57 (0.80)	47	2.25
arbiter	3.44 (0.95)	87	2.27	3.62 (0.65)	24	1.74
archaeologist	3.03 (1.07)	173	3.44	3.69 (0.78)	77	3.05
architect	2.77 (1.09)	204	3.84	3.04 (0.89)	85	3.30
artist	4.18 (0.94)	217	4.10	4.19 (0.67)	97	3.55
assistant	5.17 (1.09)	214	3.89	4.72 (0.88)	106	3.60
astronaut	2.65 (1.14)	188	3.70	2.80 (0.96)	74	2.99
astronomer	3.05 (1.13)	180	3.47	3.31 (1.01)	68	2.84
athlete	3.12 (1.02)	224	4.21	3.73 (0.71)	93	3.55
attendant	4.70 (1.22)	184	3.49	4.40 (0.80)	83	2.98
attourney	3.27 (1.00)	209	3.98	3.53 (0.75)	79	3.05
auditor	3.38 (1.10)	172	3.35	3.64 (0.80)	76	2.91
author	3.90 (0.74)	217	4.10	4.06 (0.62)	105	3.67
baker	4.73 (1.31)	219	4.09	3.81 (1.19)	104	3.67
banker	3.41 (1.10)	219	4.09	3.04 (0.94)	96	3.51

Continued on next page

(continued)

Occupation	Mean (SD)	*N*	Fam.	Mean (SD)	*N*	Fam.
barber	2.02 (1.12)	211	4.12	1.94 (0.95)	108	3.83
bartender	3.62 (1.25)	220	4.19	3.53 (1.10)	110	3.92
beautician	6.34 (0.93)	197	3.82	6.44 (0.84)	94	3.52
biologist	3.78 (0.99)	178	3.55	4.20 (0.79)	75	2.87
blogger	4.63 (1.10)	206	3.85	4.50 (0.82)	88	3.19
bookkeeper	4.71 (1.32)	199	3.73	4.39 (0.92)	97	3.36
botanist	4.40 (1.18)	149	3.11	4.41 (0.90)	64	2.62
broadcaster	3.63 (1.01)	195	3.71	3.90 (0.65)	86	3.26
broker	2.85 (1.06)	171	3.30	3.03 (0.87)	70	2.77
businessperson	3.14 (1.12)	226	4.12	3.59 (0.76)	115	3.73
butcher	1.80 (0.86)	209	3.98	1.94 (0.85)	106	3.76
carpenter	1.80 (0.90)	205	3.95	2.12 (0.97)	100	3.51
cashier	4.75 (1.11)	233	4.44	4.74 (0.94)	118	4.02
CEO	2.60 (1.10)	203	3.91	3.00 (1.07)	96	3.43
chauffeur	2.37 (1.10)	190	3.68	2.32 (0.92)	77	3.11
chef	3.35 (1.10)	219	4.24	3.18 (0.97)	114	3.96
chemist	3.16 (1.10)	173	3.44	3.86 (0.83)	99	3.46
chiropractor	3.00 (1.22)	182	3.63	3.66 (1.05)	80	3.01
choreographer	4.65 (1.44)	173	3.46	4.79 (1.06)	75	3.01
cleaner	5.11 (1.32)	218	4.05	5.35 (0.98)	114	4.06
clerk	4.41 (1.22)	205	3.88	4.51 (1.00)	92	3.27
coach	2.65 (1.16)	221	4.05	3.21 (1.01)	96	3.33
comedian	3.31 (1.00)	210	4.03	3.32 (0.85)	105	3.67
composer	3.22 (1.03)	179	3.55	3.32 (0.88)	73	2.89
computer programmer	2.74 (1.06)	214	4.03	2.99 (0.95)	100	3.57
conductor	2.56 (1.09)	181	3.50	2.99 (0.89)	73	2.87
confectioner	4.73 (1.28)	138	2.89	4.33 (0.93)	73	2.91
construction worker	1.56 (0.82)	216	4.16	1.66 (0.82)	103	3.67
cook	3.92 (1.21)	231	4.33	4.24 (1.02)	115	4.02
copywriter	3.99 (1.04)	160	3.27	4.06 (0.72)	66	2.66
counselor	4.60 (1.18)	215	3.97	4.76 (0.83)	85	3.18
curator	3.96 (1.14)	141	2.92	3.93 (0.75)	57	2.55
dancer	5.62 (1.19)	202	3.93	5.18 (1.08)	88	3.42
dental hygienist	5.38 (1.33)	219	4.14	5.39 (1.02)	108	3.77
dentist	3.34 (1.10)	228	4.32	3.71 (0.87)	116	4.10
designer	5.03 (1.21)	195	3.65	4.33 (0.96)	90	3.27
detective	2.52 (1.11)	198	3.77	3.09 (0.96)	77	3.17
developer	2.94 (1.11)	175	3.45	3.09 (1.00)	74	2.89
dietitian	5.23 (1.15)	190	3.67	5.02 (0.95)	81	3.15

Continued on next page

(continued)

Occupation	Mean (SD)	*N*	Fam.	Mean (SD)	*N*	Fam.
diplomat	2.99 (1.14)	167	3.27	3.24 (0.81)	68	2.71
director	3.12 (1.04)	188	3.64	3.34 (0.81)	93	3.39
dispatcher	4.00 (1.39)	180	3.50	3.56 (0.92)	71	2.75
doctor	3.42 (0.98)	235	4.42	3.67 (0.80)	120	4.17
dramaturg	4.02 (1.00)	42	1.57	3.94 (0.25)	16	1.47
driver	2.70 (1.10)	219	4.14	2.93 (0.88)	112	3.99
ecologist	3.75 (1.04)	147	3.07	4.05 (0.67)	56	2.50
economist	3.01 (1.10)	176	3.37	3.17 (0.84)	80	2.88
editor	3.89 (1.05)	192	3.69	3.83 (0.72)	82	3.04
educator	4.60 (1.15)	228	4.20	4.32 (0.91)	110	3.80
electrician	1.94 (0.95)	220	4.12	2.23 (0.89)	111	3.88
elementary school teacher	5.84 (1.06)	225	4.33	5.44 (1.08)	114	3.99
engineer	2.60 (1.07)	203	3.87	2.43 (1.02)	101	3.55
entrepreneur	3.28 (0.99)	214	3.95	3.36 (0.88)	91	3.33
examiner	3.53 (0.99)	150	3.07	3.82 (0.70)	83	3.03
farm manager	2.38 (1.11)	143	3.04	2.44 (0.96)	70	2.80
farmer	2.06 (0.98)	210	4.05	2.39 (0.92)	103	3.73
firefighter	1.75 (0.85)	213	4.12	2.21 (0.96)	94	3.61
fisher	2.03 (0.95)	184	3.60	2.18 (0.93)	82	3.09
flight attendant	5.86 (1.06)	210	4.07	5.36 (0.94)	106	3.83
florist	5.62 (1.19)	203	3.92	5.84 (0.97)	99	3.65
forester	2.47 (1.15)	136	2.84	2.29 (0.98)	65	2.73
freelancer	3.78 (0.80)	194	3.72	3.95 (0.64)	79	3.05
geographer	3.14 (1.09)	156	3.20	3.53 (0.80)	59	2.62
geologist	3.10 (1.20)	163	3.30	3.43 (0.81)	61	2.66
ghostwriter	3.93 (1.03)	151	3.16	3.96 (0.67)	55	2.55
goldsmith	2.45 (1.17)	121	2.75	2.97 (0.90)	70	2.77
GP	3.67 (0.98)	140	2.88	3.80 (0.82)	119	4.19
guard	2.00 (0.99)	197	3.74	2.21 (0.87)	80	2.95
guard chief	2.41 (1.15)	105	2.46	2.53 (0.97)	36	1.97
gynaecologist	4.79 (1.54)	187	3.74	4.46 (1.35)	83	3.17
hairdresser	6.00 (1.23)	217	4.20	5.75 (1.17)	119	4.22
historian	3.44 (1.14)	186	3.59	3.62 (0.92)	86	3.11
housekeeper	6.14 (1.00)	214	4.11	5.92 (0.94)	102	3.65
hygienist	5.39 (1.19)	202	3.73	5.26 (0.97)	95	3.47
illustrator	4.01 (0.93)	185	3.60	4.05 (0.77)	79	3.00
inspector	2.75 (1.09)	178	3.36	3.28 (0.80)	78	2.89
instructor	3.82 (0.94)	213	3.89	3.74 (0.69)	94	3.25
inventor	2.98 (1.07)	177	3.55	3.18 (0.90)	67	2.76

Continued on next page

(continued)

Occupation	Mean (SD)	*N*	Fam.	Mean (SD)	*N*	Fam.
investigator	2.79 (1.08)	194	3.61	3.31 (0.94)	74	2.92
janitor	2.35 (1.19)	215	4.09	2.49 (1.11)	96	3.43
journalist	4.16 (0.88)	209	3.91	3.94 (0.74)	98	3.52
judge	3.09 (1.14)	206	3.92	3.19 (0.98)	90	3.39
laborer	2.07 (1.07)	206	3.89	1.96 (0.97)	102	3.67
lawyer	3.32 (1.00)	222	4.19	3.63 (0.76)	105	3.70
librarian	5.70 (1.09)	218	4.11	5.13 (1.00)	105	3.80
life insurance agent	3.55 (1.11)	196	3.75	3.60 (0.83)	78	2.91
lifeguard	3.35 (1.18)	207	3.99	3.26 (0.97)	86	3.42
linguist	4.21 (1.00)	143	3.05	4.49 (0.67)	71	2.83
loader	1.97 (1.01)	135	2.95	2.22 (1.10)	59	2.45
machinist	2.00 (1.11)	163	3.19	3.50 (1.66)	72	2.86
magician	2.64 (1.05)	180	3.63	2.83 (0.87)	80	3.08
manager	3.43 (0.90)	230	4.25	3.65 (0.64)	116	4.07
manicurist	6.24 (1.09)	186	3.73	6.30 (0.94)	79	3.16
mason	1.80 (0.98)	148	3.12	2.05 (0.96)	61	2.54
mathematician	2.81 (1.12)	177	3.53	3.24 (0.94)	80	3.11
mayor	2.87 (1.09)	202	3.89	3.11 (1.04)	80	3.03
mechanic	1.65 (0.86)	222	4.21	1.90 (0.93)	103	3.72
mechanician	1.80 (0.96)	143	3.07	2.20 (1.03)	74	3.01
meteorologist	3.27 (1.05)	200	3.76	3.57 (0.88)	69	2.83
miller	2.32 (1.14)	94	2.31	2.61 (1.00)	51	2.37
miner	1.56 (0.85)	171	3.37	1.50 (0.76)	76	3.00
minister	2.17 (1.02)	200	3.77	3.15 (0.90)	78	2.98
model	5.80 (1.14)	207	3.98	5.21 (1.15)	91	3.41
mover	1.77 (0.99)	198	3.86	2.42 (1.18)	66	2.67
musician	3.80 (0.80)	221	4.20	3.90 (0.49)	104	3.65
neurologist	3.11 (1.13)	181	3.51	3.54 (0.86)	67	2.77
neuroscientist	3.04 (1.14)	162	3.21	3.54 (0.89)	65	2.61
neurosurgeon	2.88 (1.23)	180	3.47	3.31 (1.00)	68	2.84
newscaster	3.93 (0.97)	218	4.01	4.04 (0.57)	94	3.48
notary	4.34 (1.18)	172	3.43	3.79 (0.97)	48	2.23
nun	6.94 (0.23)	178	3.70	6.95 (0.22)	82	3.13
nurse	6.02 (0.99)	230	4.35	5.74 (0.96)	120	4.20
nutritionist	5.07 (1.18)	200	3.79	4.83 (0.92)	92	3.30
obstetrician	4.47 (1.37)	158	3.25	4.27 (1.27)	56	2.49
occupational therapist	4.48 (1.22)	186	3.49	4.91 (0.84)	85	3.11
officer	2.60 (1.11)	218	4.01	3.38 (0.90)	88	3.17
optician	3.71 (1.06)	191	3.63	4.09 (0.69)	108	3.84

Continued on next page

(continued)

Occupation	Mean (SD)	*N*	Fam.	Mean (SD)	*N*	Fam.
painter	3.12 (1.30)	216	4.05	3.27 (1.05)	106	3.75
paralegal	4.86 (1.15)	180	3.44	4.26 (0.92)	72	2.75
paramedic	3.37 (1.09)	211	3.98	3.92 (0.72)	100	3.63
pathologist	3.58 (1.07)	151	3.05	3.79 (0.79)	58	2.48
pawnbroker	2.44 (1.09)	147	3.17	2.84 (0.84)	68	2.74
pediatrician	4.39 (1.20)	209	3.91	4.49 (1.01)	80	3.14
personnel director	3.99 (1.09)	158	3.14	4.32 (0.96)	72	2.80
pharmacist	3.92 (1.19)	222	4.16	4.31 (0.85)	108	3.76
photographer	4.03 (1.08)	223	4.16	3.79 (0.80)	100	3.69
physician	3.40 (1.08)	229	4.26	3.65 (0.86)	99	3.42
physicist	3.03 (1.12)	175	3.38	3.46 (0.96)	72	2.86
pilot	2.34 (1.03)	211	4.03	2.78 (0.91)	94	3.55
planner	4.67 (1.43)	169	3.32	3.87 (0.92)	71	2.83
plumber	1.70 (0.85)	213	4.09	2.05 (0.94)	111	3.89
poet	4.07 (1.11)	188	3.69	3.94 (0.82)	77	2.98
politician	3.07 (1.02)	220	4.15	3.40 (0.75)	97	3.59
pope	1.12 (0.33)	186	3.77	1.06 (0.23)	72	3.11
potter	3.86 (1.32)	135	2.93	4.18 (0.88)	68	2.74
practitioner	3.84 (1.00)	170	3.29	3.94 (0.61)	71	2.81
priest	1.40 (0.76)	195	3.87	1.76 (0.94)	80	3.23
producer	3.21 (1.00)	172	3.46	3.45 (0.89)	74	2.89
professor	3.27 (1.08)	220	4.17	3.28 (0.92)	103	3.67
programmer	2.83 (1.14)	202	3.85	3.16 (1.07)	86	3.26
prosecutor	2.96 (1.10)	206	3.75	3.30 (0.89)	76	2.86
psychiatrist	3.95 (1.00)	207	3.90	4.04 (0.75)	85	3.14
psychologist	4.26 (1.17)	219	3.96	4.39 (0.83)	77	3.08
public relations director	4.31 (1.32)	156	3.11	4.13 (0.96)	61	2.59
publicist	4.41 (1.17)	160	3.25	4.32 (0.77)	71	2.69
radiologist	3.60 (1.14)	181	3.47	4.12 (0.79)	77	3.05
realtor	4.42 (1.19)	212	4.02	4.18 (0.85)	71	2.87
receptionist	6.06 (1.01)	227	4.28	5.59 (1.03)	120	4.17
registered nurse	6.11 (0.95)	221	4.25	5.79 (1.00)	117	4.01
repairer	2.26 (1.08)	198	3.72	2.78 (0.97)	79	2.98
researcher	3.62 (0.97)	203	3.78	3.94 (0.67)	86	3.17
roofer	1.54 (0.86)	200	3.86	1.64 (0.84)	91	3.41
salesperson	3.75 (0.94)	232	4.33	3.87 (0.87)	117	4.00
school administrator	4.29 (1.31)	207	3.80	5.17 (1.18)	102	3.63
scientist	3.13 (1.04)	205	3.90	3.60 (0.82)	97	3.40
screenwriter	3.62 (0.95)	176	3.43	3.81 (0.64)	73	2.82

Continued on next page

(continued)

Occupation	Mean (SD)	*N*	Fam.	Mean (SD)	*N*	Fam.
secretary	6.11 (1.14)	221	4.24	5.92 (0.88)	114	4.05
servant	4.32 (1.13)	160	3.28	4.10 (0.96)	71	2.88
sewer	4.81 (2.04)	134	2.91	4.73 (1.90)	67	2.74
sheriff	1.97 (1.03)	208	4.00	2.12 (0.98)	66	2.73
shoemaker	2.54 (1.07)	145	3.18	2.67 (0.91)	73	2.94
singer	4.53 (0.94)	221	4.20	4.30 (0.59)	105	3.74
social worker	5.34 (1.34)	199	3.84	5.20 (1.02)	96	3.46
sociologist	4.54 (1.13)	163	3.22	4.40 (0.91)	67	2.71
soldier	1.88 (0.99)	212	4.05	1.99 (0.96)	101	3.70
specialist	3.59 (0.85)	170	3.32	3.85 (0.63)	74	2.73
speechwriter	3.86 (0.95)	162	3.27	3.70 (0.71)	63	2.59
statistician	3.15 (1.03)	158	3.12	3.39 (0.90)	72	2.84
stockbroker	2.48 (1.08)	183	3.48	2.52 (0.91)	75	2.98
supervisor	3.41 (1.00)	229	4.19	3.85 (0.64)	113	3.73
surgeon	2.88 (1.11)	208	3.97	3.12 (0.97)	95	3.48
tailor	3.75 (1.48)	180	3.55	3.05 (1.28)	81	3.15
tattooist	3.09 (1.09)	197	3.77	3.42 (0.86)	89	3.33
teacher	5.14 (1.12)	236	4.48	4.63 (0.85)	127	4.38
technician	2.77 (1.12)	195	3.62	3.20 (0.92)	92	3.24
telemarketer	4.01 (1.15)	201	3.86	4.11 (0.72)	79	3.05
theologian	2.78 (1.12)	133	2.83	3.43 (0.87)	44	2.23
therapist	4.78 (1.18)	209	3.99	4.74 (0.87)	91	3.31
umpire	1.80 (0.94)	173	3.56	2.68 (1.06)	74	3.01
usher	3.14 (1.38)	171	3.41	3.62 (1.37)	69	2.72
veterinarian	4.05 (1.22)	217	4.08	4.26 (0.92)	103	3.70
waiter	3.33 (1.60)	230	4.37	3.50 (1.23)	119	4.14
welder	1.72 (0.97)	186	3.64	1.79 (0.86)	77	3.06
woodworker	1.99 (1.01)	185	3.60	2.36 (0.97)	83	3.17
worker	3.52 (0.91)	222	4.14	3.72 (0.67)	117	3.98
writer	4.09 (0.81)	213	4.01	4.01 (0.54)	102	3.57
x-ray technician	3.99 (1.24)	198	3.82	4.20 (0.81)	82	3.10

Appendix B

Appendices for Chapter 5

B.1 Appendix: Prompts used to generate Prime pictures

Click here to return to Chapter 5.

We used the following prompts on Midjourney to generate our pictures of the different occupations and genders. We order from most to least "feminine" (as per our study where we asked participants to rate occupations as stereotypically masculine or feminine; Chapter 4).

Several differences between female and male prompts stem from small problems we encountered when applying the same prompt across genders. For example, when generating pictures of nurses, the generated female nurses looked substantially younger, which is why we ended up using different ages (40 vs 35) in the prompts for the two genders. None of these prompts immediately resulted in our final picture. Instead, upon finding a sufficient prompt, we often had to run a prompt several times, had to generate variations of the first near-acceptable picture, and had to instruct Midjourney to revise sections of a picture (e.g., because something illogical was depicted that outed the picture as AI-generated).

- *Secretary*
  - **Female:** 40-years-old female secretary stands and takes notes in a notebook, she has a neutral facial expression, plane white background, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –ar 1:1
  - **Male:** 40-years-old male secretary stands and takes notes in a black notebook, he wears a light-grey button-up shit and has an attentive

and neutral facial expression, plane white background, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –ar 1:1 –v 5.2

- *Nurse*
  - **Female:** Portrait of a 40-years-old female nurse, she has a professional demeanor, plane white background, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –ar 1:1
  - **Male:** Portrait of a 35-years-old male nurse with lightblue scrubs and stethoscope, he has a professional demeanor, plane white background, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –ar 1:1
- *Flight attendant*
  - **Female:** female flight attendant in uniform with a small scarf and a hat, she has a neutral facial expression, plane white background, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –ar 1:1 –v 5.2
  - **Male:** male flight attendant with a blue vest and a light-blue tie, the vest has a golden airline pin bage, the flight attendant has a neutral facial expression, plane white background, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –v 5.2 –ar 1:1
- *Florist*
  - **Female:** Portrait of a 40-years-old female florist, she has a professional demeanor, plain white background, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –v 6.0 –ar 1:1
  - **Male:** Portrait of a 35-years-old male florist, he has a professional demeanor, plain white background, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –v 6.0 –ar 1:1
- *Cleaner*
  - **Female:** Portrait of a Caucasian 40-years-old female cleaner, she holds a cleaning product, she has a neutral facial expression, plane white background, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –ar 1:1

  - **Male:** Portrait of a Caucasian 40-years-old male cleaner, he holds a cleaning product, he has a neutral facial expression, plane white background, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –ar 1:1

- *Surgeon*

  - **Female:** 40-years-old female surgeon holds a scalpel, she wears a surgical cap and a surgical mask, plane white background, she has a professional demeanor, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –ar 1:1
  - **Male:** [name of the file of the female surgeon picture] as a male surgeon –iw 2 –ar 1:1

- *Farmer*

  - **Female:** Portrait of a 40-years-old female farmer in blue overalls and red plaid shirt wearing a cap, she is holding a chicken and has a neutral facial expression, plane white background, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –ar 1:1
  - **Male:** Portrait of a 40-years-old male farmer in blue overalls and red plaid shirt wearing a cap, he is holding a chicken and has a neutral facial expression, plane white background, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –v 5.2 –ar 1:1

- *Soldier*

  - **Female:** Portrait of a 40-years-old female soldier, she wears a uniform, she has a professional demeanor, plane white background, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –ar 1:1
  - **Male:** Portrait of a 35-years-old male soldier, he has a professional demeanor and slightly ruffled hair, plane white background, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –ar 1:1

- *Firefighter*

  - **Female:** Portrait of a 35-years-old female firefighter, she wears a black and neon yellow firefighter uniform and a helmet, she has a professional

demeanor, plane white background, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –v 6.0 –ar 1:1

  - **Male:** Portrait of a 40-years-old male firefighter, he wears a uniform, he has a professional demeanor, plane white background, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –v 6.0 –ar 1:1

- *Construction worker*

  - **Female:** Portrait of a 40-years-old female construction worker, she wears a neon orange safety vest and a helmet, she has a professional demeanor, she stands in front of a plain plain white wall, she has brown eyes, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –ar 1:1
  - **Male:** Portrait of a 40-years-old male construction worker, he wears a neon orange safety vest and a helmet, he has a professional demeanor, he stands in front of a plain plain white wall, he has brown eyes, realistic documentary photography taken with a cinematic 35-millimeter lens, F 1.8, ultra-detailed and intricate photo –ar 1:1

B.2 Appendix: List of target words and their origin

Click here to return to Chapter 5.

Note: Abele et al. (2008) and Abele et al. (2016) have a lot of overlap, so most adjectives sourced here from the former could have also been sourced from the latter.

Agency adjectives

- **Abele et al. (2008):** “able”, “active”, “assertive”, “capable”, “creative”, “determined”, “independent”, “intelligent”, “persistent”, “rational”

Communion adjectives

- **Abele et al. (2008):** “altruistic”, “considerate”, “friendly”, “helpful”, “loyal”, “polite”, “sensitive”, “supportive”, “sympathetic”, “trustworthy”

Agency nouns

- **Derived from adjectives of Abele et al. (2016):** "cleverness" [from "clever"]
- **Derived from adjectives of Abele et al. (2008):** "ambition" [from "ambitious"], "confidence" [from "self-confident"], "industriousness" [from "industrious"],
- **Pietraszkiewicz et al. (2019):** "decision", "skill", "willpower"
- **Trapnell and Paulhus (2012):** "autonomy", "competence" (notably, Abele et al., 2008, also has "competent" as an agency-related adjective), "status"

Communion nouns

- **Derived from adjectives of Abele et al. (2016):** "empathy" [from "empathetic"]
- **Derived from adjectives of Abele et al. (2008):** "affection" [from "affectionate"], "fairness" [from "fair"], "morality" [from "moral"], "openness" [from "open"], "tolerance" [from "tolerant"]
- **Trapnell and Paulhus (2012):** "forgiveness", "harmony", "honesty" (notably, Abele et al., 2008, also has "honest" as a communion-related adjective), "humility"

B.3 Appendix: Curation of word lists

Here we describe how we sampled and manually revised our list of distractor words and practice trial words. As with our analysis scripts and stimuli, the script we used for sampling and revising these lists is available at https://osf.io/frq8c/overview?view_only=5ed5db2669b24b4e8e682b338b0ea382.

Curation of a large candidate list To create suitable lists of distractor words as well as words that we could use in the practice trials for our experimental task (see Section 5.2.4), we assessed the psycholinguistic properties of our list of target words with the South Carolina Psycholinguistic Metabase (SCOPE; Gao et al., 2023), a database that combines the psycholinguistic data from several different psycholinguistic databases. They list several different variants of similar constructs, for example several different ways in which people's familiarity with a word was assessed. We searched the SCOPE database for those variants of important psycholinguistic properties that had few missing values for our list

of stereotype-related target words. Subsequently, we tried to align our list of distractor words and practice words on these (variants of the) properties.

Specifically, from our list of candidate words for the distractor words and the practice trial words, we excluded all words that were not highly familiar as assessed by the proportion of people who knew the word in Brysbaert et al. (2014)'s sample. We retained only words that 96% or more of people knew, as that was the lowest observed value amongst our stereotype-relevant target words. Additionally, we excluded all words that were either more positively valenced than the most positively valenced target word or more negatively valenced than the most negatively valenced target word, as assessed by Warriner et al. (2013) with a Likert rating of how pleasant a word is (from "unhappy" to "happy"). We further excluded from our list of candidate words every word with a notable gender association (i.e., a value of < 3 or > 5 on the 7-point Likert scale of Scott et al., 2019, which goes from "feminine" to "masculine"). Finally, we excluded all adjectives that ended on "-ing", as they can also be used as nouns (e.g. as in "The tiring of people has become a problem."), making the word categorization task ambiguous.

Revising the final lists After reducing our list of candidate words in that manner, we randomly sampled sets of 40 adjectives and 40 nouns that fulfilled two conditions: Firstly, they had to be approximately similarly long as our target words. Initially we sampled them to be within 0.75 letters of the average length of our target words, though later we manually revised the word list until the difference was down to 0.1 letters. Secondly, upon recommendation of Gao et al. (2023), we tried to control for the "Association Frequency 123" – a proxy for how semantically embedded a word is in memory, as assessed by the count of how many times a word was amongst the first three words people brought up in a free association task (De Deyne et al., 2019). Specifically, we sampled in a way that ensured that our distractor nouns had a mean log(AssociationFrequency123) that is close to the mean of our target adjectives, and that our distractor adjectives had a mean close to that of our target nouns. We did so because our target adjectives have slightly higher values than our target nouns and we wanted to prevent participants from developing heuristics like "if I can recognize it faster, it's an adjective". We sampled based on the natural logarithm of Association Frequency rather than based on its raw scores, as the raw counts were highly skewed (i.e., some words had counts in the thousands which highly influenced the mean of our words).

After sampling and skimming multiple such lists, we revised two lists that would become our practice trial words and our distractor words, by manually and iteratively removing words and, in their stead, sampling new ones which still adhered to the word length and Association Frequency criteria.

We resampled words for the following reasons:

- They were occupational terms, which could have interfered with how our subsequent occupation prime pictures are processed.
- They were gender-related terms (e.g., bikini) or terms related to gender roles or stereotypes (e.g., baby), which could have led to unintended priming effects that carry over into subsequent hypothesis-relevant trials.
- They described people's physical appearance, which could have influenced how participants process the person on the subsequent prime picture.
- They related to something that is depicted on a prime picture (e.g., the name of a flower, since flowers are depicted on the florists' prime pictures), as that could have influenced the processing of a subsequent prime.
- They are related to agency or communion (e.g., "adventurous"), which could have led to unintended priming effects on the subsequent trial.
- They could be interpreted either as adjectives or nouns (e.g., "visual"), making the categorization task ambiguous.

B.4 Appendix: Preregistered analyses we forewent

There are two sets of analyses we forewent, compared to what we preregistered.

Click here to return to Chapter 5.

Bayesian DDMs We preregistered that we would also fit a series of Bayesian hierarchical Diffusion Decision Models to evaluate whether priming took place (DDM; Ratcliff et al., 2016; previously applied to sequential priming data by Tsamadi et al., 2020; Voss et al., 2013). We ultimately decided against fitting these models, as that would have taken months but was unlikely to lead to additional insights, given that we found no effects for reaction time or accuracy (e.g., Wilson & Collins, 2019, argue against fitting cognitive models, if no statistical effects can be found with inferential statistical models).

(Generalized) Linear mixed models We preregistered that we would run all of our (generalized) linear mixed models with factor-coded in addition to the regular contrast-coded predictors. That was a mistake. In addition to complicating the random effects structure of our model (i.e., R struggles with factor-coded random slopes; see e.g., Bates et al., 2015), this would have also not provided additional information about our fixed effects structure, as our fixed effect factors were binary (i.e., for factors with two levels, it does not matter whether R treats them as a numeric or a factor).

B.5 Appendix: Simplest valid analysis

Click here to return to Chapter 5.

The simplest analysis one could perform to test whether there are effects of priming on participants' accuracy or reaction time is to run a logistic regression on whether or not they answered correctly and a linear regression on their (raw) reaction times. The model formulas for both models are $DependentVariable \sim Congruency+BlockType:Congruency+BlockType+LexicalCategory+Replicate$. That means we included two predictors that could indicate that priming took place – the main effect of Congruency and the interaction effect between Congruency and BlockType – as well as three nuisance factors.

For accuracy, participants performed worse on nouns, $\beta_{LexicalCategory} = -1.445$, $z = -33.567$, $p < .001$ and a bit better the more often they had seen a target word, $\beta_{Replicate} = 0.036$, $z = 2.007$, $p = .045$. There is no evidence for priming, as the main effect of congruency is insignificant and trends in the wrong direction, $\beta_{Incongruent} = 0.057$, $z = 1.020$, $p = .308$. Similarly the interaction effect is insignificant and trends in the direction of "more priming if people focus on the shape" which is inconsistent with the hypothesis, $\beta_{Shape*Incongruent} = -0.098$, $z = -1.234$, $p = .217$.

For reaction times, participants were slower on blocks that focused on the shape, $\beta_{BlockType} = 0.0983$, $t = 3.092$, $p = .002$, and became faster the more often they had seen a target word, $\beta_{Replicate} = -0.196$, $t = -19.514$, $p < .001$. There is no evidence for priming, as the parameter estimate for the congruency main effect was virtually zero (and trended in the wrong direction), $\beta_{Congruency} = -0.001$, $t = 0.032$, $p = .978$. The interaction, too, is insignificant and trends in the hypothesis-irrelevant direction of "on blocks with a shape focus, there is counter-priming" (i.e., faster responding on incongruent trials), $\beta_{Shape*Incongruent} = -0.031$, $t = -0.683$, $p = .495$.

The simplest valid Bayesian test of the hypothesis, a simple linear regression with the same predictors as our Frequentist linear regression and Jeffreys-Zellner-Siow (JZS) prior distributions, allows us to assess whether we simply do not know whether Congruency and its interaction with BlockType are relevant – *absence of evidence* – or whether we have evidence against them being relevant to explain our data – *evidence for absence*. To that end we fit the full regression, $Model_1$, then a variant without the interaction effect, $Model_2$, and finally a model without the interaction effect and the Congruency main effect, $Model_3$. We can then simply compute ratios of how well the data is described by one model vs the other, so-called *Bayes Factors*.

The data are approximately 35.8x more probable under a model without the interaction than under a model that assumes that an interaction exists, $BF_{M2}/BF_{M1} = 35.758$. The data is even more probable under a model that

also assumes that no Congruency main effect exists, $BF_{M3}/BF_{M2} = 37.603$. At least as evaluated by its (potential) impact on reaction time, we obtained strong evidence against priming having taken place in our experiment.

Appendix C

Appendices for Chapter 6

C.1 A brief and oversimplified introduction to Structural Equation Modeling and Measurement Invariance

Click here to return to Chapter 6, or here to return to Chapter 7.

Structural Equation Modeling *Structural Equation Modeling* is a class of statistical methods for translating the relationships one assumes to exist between unobservable "latent" variables – and between those variables and observable indicators thereof – into mathematical predictions. This involves at least three steps: specifying which relationships one assumes to exist, fitting the model algorithmically (i.e., finding values for the model's parameters that, given the specified relationships, best explain the observed patterns of variances and covariances of test-takers' responses to different test items), and evaluating the fit (i.e., assessing whether model-predicted patterns of variances and covariances are sufficiently similar to patterns observed in actual data).

Figure C.1 depicts an example model. Circles represent unobservable latent variables, rectangles observed scores – how test-takers answered to six different questionnaire items (A-F) – and arrows the relationships the modeler assumed. These relationships can be either positive (e.g., the more a respondent possesses of a latent variable, the more likely they are to endorse the item) or negative (e.g., the more a respondent possesses of a latent variable, the less likely they are to endorse it). The modeler specified that people's responses to items A, B, and C are all influenced by the same underlying latent variable (i.e., the big orange circle). For example, if these three questions are about people's views on immigration, we assume that people's overall support or opposition to immigration influences

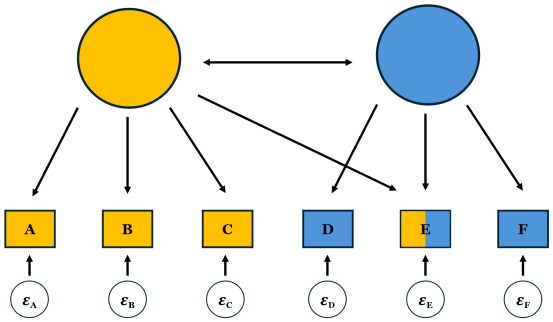


Figure C.1: Example of a structural equation model with two (substantive) latent variables.

their responses to these questions. Additionally, responses to each particular item are also caused by other unobservable influences (the small white circles). These can include random measurement error (e.g., stemming from misclicks), but also the impact of latent variables that one did not formally include in the model (e.g., on a question about immigration from Germany, people's opinions on Germans can be influential), or the impact of other sources of systematic error (e.g., items with double negatives, like "Immigrants should not be denied entry.", could lead to frequent misunderstandings).

By specifying relationships among latent variables and indicators, the modeler implicitly specifies which patterns of correlations should arise between test-takers' responses. For example, here we model the existence of two latent variables (e.g., opposition to immigration, orange, and opposition to federal spending, blue) which we assume to be related. All else being equal, this implies that the correlation between responses that test-takers give to items A and C should be stronger than the correlation between responses to A and F: Items A and C are both related to the same latent variable (e.g., both assess immigration), while items A and F relate only indirectly through the relationship of their latent variables (e.g., item A is related to immigration, views on immigration are related to views on federal spending, and item F is related to federal spending).

SEM can model several different types of relationships between items and/or latent variables which all have implications for what kind of correlations should be observed. For example, responses to some items, like E, are influenced by multiple latent variables (e.g., test-takers' responses to "The Netherlands should spend more money on immigrants." is associated both with their views on immigration and their views on federal spending) – implying that item A should have a stronger correlation with E than with D or F, given that the items share a latent variable. Alternative models, for example that responses to all six items instead reflect one substantive latent variable (e.g., political conservatism), would imply very different correlational patterns. A comparison between model-implied and actually observed patterns of correlations among test-takers' responses enables

one to test whether a theoretical model is (reasonably) compatible with the data.

Measurement invariance Measurement invariance (or "measurement equivalence") refers to the degree to which a measurement tool assesses a construct in the same way – across different groups of test-takers (e.g., humans vs silicon samples), different measurement contexts, or different points in time. In the context of Structural Equation Models (SEMs), one, in essence, tests to what extent two groups of responses for the same measurement tool (e.g., responses from two groups of test-takers, or from the same group at two different times) can be described by a SEM with the same parameters. This is a matter of degrees: The more stages a group comparison passes, the more similar the two groups of responses are (in our case: the more similar silicon and human test-takers are in how they use the questionnaire). Here, I discuss only the two most basic (and hence most foundational) stages. As I will describe below, failing either of these stages would indicate that (current) LLM imitators are limited in their ability to reproduce the basic structure of human questionnaire responses.

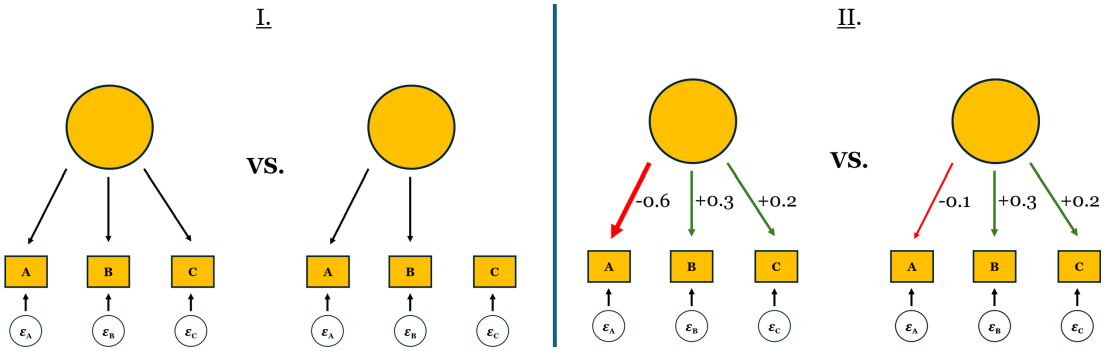


Figure C.2: Example violations of *Configural* (I.) and *Metric* (II.) *Measurement Invariance*.

Configural invariance requires that, for the two groups of compared responses, the same relationships exist between latent variables and their indicators (i.e., both SEMs have the same arrows from circles to rectangles). Figure C.2.I violates this invariance: For the left group, item C is related to the latent variable (e.g., for those test-takers, responses to C correlate with responses to A and B) but for the right group, it does not. If configural invariance does not hold between human respondents and their LLM imitators, the LLMs did not process questionnaire items in a manner that resembles human responding even at the most basic structural level (e.g., it did not "understand" that all items A-C are about immigration). In such cases, comparing responses between the two types of test-takers is (virtually) meaningless.

Metric (or *weak*) *invariance* requires that, for the two compared groups of

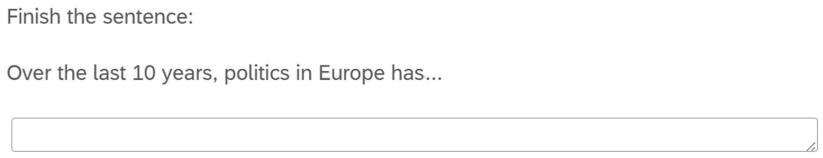
responses, relationships between latent variables and their indicators (i.e., "factor loadings") are (approximately) equal. Figure C.2.II violates this invariance: Item A is strongly related to the latent variable for the left group, while the relationship is weak for the right group. If metric invariance does not hold between responses of humans and their LLM imitators, LLMs do not accurately reproduce how differences in the underlying construct should be reflected in responses to different questionnaire items. In more severe violations, not only the magnitude (as in Figure 2.II) but also the directionality can be wrong: An LLM imitator might, for example, endorse an item to indicate support for immigration, even though (for humans) endorsement of that item actually indicates opposition to immigration.

C.2 Example block

Click here to return to Chapter 6.

Below, the questionnaire items from the block (purportedly) about international events and international politics. Italicized words indicate comments on the item (i.e., not text that is visible to respondents). Distractor items were designed to be plausible questionnaire items with varying response formats.

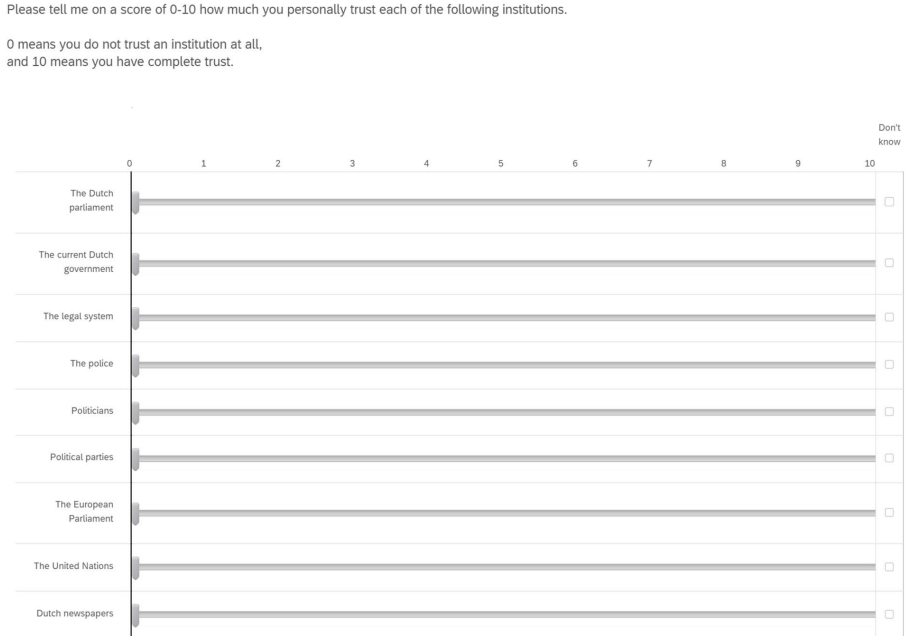
Distractor (open text response)



Distractor (4 non-symmetrical response categories)



European Social Survey: "Trust in Institutions" scale (original response format and introductory text). *One additional item was included – "the current Dutch government" – whose responses should correlate with Mansoor's "trust on government" scale.*



Distractor (4 non-symmetrical response categories)

How much interest do you have in hearing news about the following foreign countries?

	Not at all	Little	Somewhat	A great deal
Belgium	○	○	○	○
Germany	○	○	○	○
the UK	○	○	○	○
Spain	○	○	○	○
France	○	○	○	○
the USA	○	○	○	○
Turkey	○	○	○	○
Israel	○	○	○	○
Ukraine	○	○	○	○
Russia	○	○	○	○
China	○	○	○	○
India	○	○	○	○
Indonesia	○	○	○	○
Marocco	○	○	○	○
South Africa	○	○	○	○

***Distractor (3 symmetrical response categories and an "unsure" option)**. Countries are identical to the ones from the question before.*

How do you think the financial situation of these foreign countries will develop, in the next decade?

	Will become less prosperous	Will be (about) equally prosperous	Will become more prosperous	Unsure
Belgium	○	○	○	○
Germany	○	○	○	○
the UK	○	○	○	○
Spain	○	○	○	○
France	○	○	○	○
the USA	○	○	○	○
Turkey	○	○	○	○
Israel	○	○	○	○
Ukraine	○	○	○	○
Russia	○	○	○	○
China	○	○	○	○
India	○	○	○	○
Indonesia	○	○	○	○
Marocco	○	○	○	○
South Africa	○	○	○	○

Distractor (4 symmetrical response categories). *The location of this question type (i.e., at the end of the block), the introductory text, and answer format were identical for each block. Across all blocks, about half of the news items (but in this case only the 3rd, 4th, and 7th) were fictitious.*

Please indicate whether you believe a certain news event occurred and how certain you are about this.

	I am certain this happened	I think this happened but I am not certain	I don't think this happened but I am not certain	I am certain this did not happen
German Chancellor Olaf Scholz lost a "vote of no confidence", paving the way for a federal election in 2025.	○	○	○	○
President of South Korea Yoon Suk Yeol declared martial law - a decision that was reversed by the South Korean parliament within 6 hours.	○	○	○	○
Following Xi Jinping's surprise retirement announcement, the Chinese Communist Party (CCP) announced that Vice President Han Zheng will succeed him in 2026.	○	○	○	○
Sweden withdrew its veto, paving the way for Turkey's ascension to NATO by the end of 2025.	○	○	○	○
Former president Jimmy Carter died, aged 100.	○	○	○	○
Romania's constitutional court annulled the second round of presidential elections in Romania.	○	○	○	○
Negotiations about a national union of Australia and New Zealand collapsed after the first rounds of talks.	○	○	○	○
Donald Trump won the 2024 election for US president by defeating Democratic candidate Kamala Harris.	○	○	○	○
Israel performed multiple airstrikes on military targets in Syria, ostensibly to prevent Syria's new provisional government from acquiring the there-stored weapons.	○	○	○	○
For the first time in South Africa's (democratic) history, the ANC party failed to win a majority of the vote in a general election.	○	○	○	○

C.3 Appendix: Questions & Response labels

Click here to return to Chapter 6.

Question content	**What participants evaluate**	**(English translations of the) Response options**
Gender		"Male", "Female", "Other; please specify below", "Prefer not to say"
Country of residency		"The Netherlands", "Others"
National papers they read last week		Participants could choose any number of the following: "De Telegraaf", "Algemeen Dagblad", "De Volkskrant", "NRC", "Trouw", "Het Financieele Dagblad", "Reformatorisch Dagblad", "Nederlands Dagblad", "NOS.nl", "nu.nl", "Metro Nieuws", "None of these newspapers"
Regional papers they read last week		Participants could choose any number of the following: "De Limburger", "De Gelderlander", "De Stentor", "Noordhollands Dagblad", "Dagblad van het Noorden", "Brabants Dagblad", "Tubantia", "Eindhovens Dagblad", "BN De Stem", "Leeuwarder Courant", "None of these newspapers"

Question content	**What participants evaluate**	**(English translations of the) Response options**
From which sources they receive news		Participants could choose any number of the following: "TV", "Radio", "Friends or Family", "Newspapers", "Social media", "Other"
Which of the 4 biggest newspapers they know well enough to evaluate		Participants could choose any number of the following: "De Telegraaf", "Algemeen Dagblad", "De Volkskrant", "NRC"
Quality of newspapers	[all newspapers they previously reported reading or knowing]	"Unsure / no opinion", "Very poor", "Poor", "Average", "Good", "Very good"
Political bias of newspapers	[all newspapers they previously reported reading or knowing]	"Unsure / no opinion", "Strongly biased in favor of the political left", "Biased in favor of the political left", "Somewhat biased in favor of the political left", "Unbiased / Not biased in favor of either political direction", "Somewhat biased in favor of the political right", "Biased in favor of the political right", "Strongly biased in favor of the political right"
Accuracy and Factuality of newspapers	[all newspapers they previously reported reading or knowing]	"Unsure / no opinion", "Very inaccurate", "Inaccurate", "Average", "Accurate", "Very accurate"

Question content	**What participants evaluate**	**(English translations of the) Response options**
Reporting focus of newspapers	[all newspapers they previously reported reading or knowing]	"Unsure / no opinion", "Strongly focuses too much on international news", "Focuses too much on international news", "Focuses slightly too much on international news", "Strikes a perfect balance between regional and international news", "Focuses slightly too much on regional news", "Focuses too much on regional news", "Strongly focuses too much on regional news"
Tone of newspapers	[all newspapers they previously reported reading or knowing]	"Unsure / no opinion", "Very sensationalist", "Sensationalist", "Somewhat sensationalist", "Somewhat neutral in tone", "Neutral in tone", "Very neutral in tone"

Question content	**What participants evaluate**	**(English translations of the) Response options**
Topics they find interesting	"Sport", "Culture", "Movies and/or TV shows", "Boardgames", "Videogames", "Fashion", "Celebrities", "Politics", "Local events", "Domestic (Dutch) events", "International events", "Science", "Technology", "Education", "Philosophy", "Healthy living", "Cooking", "Economy", "Stock market"	"Very interested", "Quite interested", "Hardly interested", "Not at all interested"
Place of birth		"Born in the Netherlands", "Born in a non-European country that is part of the kingdom of the Netherlands (e.g., in Curacao, Saba, etc.)", "Born somewhere else", "Born in Belgium"

Question content	**What participants evaluate**	**(English translations of the) Response options**
Job sector		"I am currently unemployed", "I am a student", "Healthcare", "Information Technology (IT)", "Real Estate", "Finance", "Retail", "Transportation", "Entertainment", "Arts", "Gastronomy (Food Services)", "Public Services (e.g., public transport, government)", "Education and Childcare", "Government", "Manufacturing", "Construction", "Agriculture", "Tourism and Hospitality", "Legal Services", "Professional Services (e.g., consulting, marketing, accounting)", "Research", "Professional experiment participant", "Other"
Relationship status		"Married", "In a long-term relationship (5+ years)", "In a relationship (< 5 years)", "Widowed", "Single"
Parenthood		"No", "Yes, one child and not planning on additional children", "Yes, one child and planning on additional children", "Yes, more than one child"

Question content	**What participants evaluate**	**(English translations of the) Response options**
Household income		"less than 20000 euros (less than 1543 per month)", "between 20000 and 40000 euros (between 1543 and 3086 per month)", "between 40000 and 60000 euros (between 3086 and 4630 per month)", "between 60000 and 80000 euros (between 4630 and 6173 per month)", "between 80000 and 100000 euros (between 6173 and 7716 per month)", "more than 100000 euros (more than 7716 per month)"
Residential environment		"Small village (less than 3,000 people)", "Village or Small Town (3,000 to 10,000 people)", "Medium-sized Town (10,000–50,000 people) that is not a suburb of a big city", "Small city (50,000–100,000 people) that is not a suburb of a big city", "Medium-sized city that is not a suburb of a big city (100,000–300,000 people)", "Suburb of a big city (e.g., Amstelveen)", "Large city (more than 300,000 people)"

Question content	**What participants evaluate**	**(English translations of the) Response options**
Highest attained educational degree		"Primary school", "Highschool (VMBO, HAVO, VWO)", "Meester" / Master craftsperson, "MBO", "Bachelor's degree (HBO, WO)", "Master's degree (HBO, WO)", "PhD or MD"
Political left-right self-placement (question from the economic left-right scale of the EES; Schmitt et al., 2016)		Integer value between 0 and 10, where 0 represents "left" and 10 represents "right"; participants could also refuse to answer
Subjective happiness (European Social Survey European Research Infrastructure, 2024)		Integer value between 0 and 10, where 0 represents "Very unhappy" and 10 represents "Very happy"; participants could also refuse to answer
Extent to which activities are "real sport"s	"Car racing (e.g., Formula 1)", "Chess", "Billiards", "ESports (competitive video gaming)", "Eating competitions"	Integer value between 0 and 10, where 0 represents "Clearly not a sport" and 10 represents "Clearly a sport"

Question content	**What participants evaluate**	**(English translations of the) Response options**
Agreement with individual sport-related statements	"Practicing team sports is less valuable than practicing individual sports", "In real sports, you move both your legs and your arms", "For good health, exercising is more important than healthy eating", "People benefit from having a sports team to cheer for", "Using performance-enhancing drugs is worse than acting against the core values of sportsmanship", "Admiring athletes for their skills is silly", "Professional sports have become far too commercial"	"Fully agree", "Agree", "Disagree", "Fully Disagree", "Unsure"
Whom they trust to advocate for changes to a sport that are healthy for the sport	"Amateur athletes", "Professional athletes", "Coaches", "Individual fans of a sports club", "Voting members of a sports club"	"Fully trust", "Trust somewhat", "Distrust somewhat", "Distrust", "Unsure"

Question content	**What participants evaluate**	**(English translations of the) Response options**
Agreement with individual culture and entertainment-related statements	"Listening to an audiobook is just as valuable as reading the book", "Our entertainment industry is too Americanized", "A well-rounded person has many different hobbies", "Poetry is overrated", "Art exhibitions deserve more funding", "Hobbies were invented to distract us from what really matters", "Ordinary people entertaining each other via the internet will replace most forms of traditional entertainment", "Hiphop is music", "Television channels have become redundant", "Making people laugh is the best form of entertainment", "Culture reached its peak in the 20th century"	"Agree", "Neither agree nor disagree", "Disagree"

Question content	**What participants evaluate**	**(English translations of the) Response options**
Perceived motives of artists and entertainers	"Actors", "Athletes", "Drawing teachers", "Music teachers", "Novel writers", "Orchestra musicians", "Pop musicians", "Social media personalities (e.g., YouTubers)", "Zookeepers"	Integer value between 0 and 10, where 0 represents "Solely interested in benefitting themselves" and 10 represents "Solely interested in bringing people joy"
Views on Immigration (European Social Survey European Research Infrastructure, 2024)	"People of the same skin color or ethnic group as most Dutch people", "People of a different skin color or ethnic group than most Dutch people", "People from poorer countries outside Europe"	"Allow many to come and live here", "Allow some", "Allow a few", "Allow none"
Perceived economic effect of immigration (European Social Survey European Research Infrastructure, 2024)		Integer value between 0 and 10, where 0 represents "Bad for the economy" and 10 represents "Good for the economy"
Perceived cultural effect of immigration (European Social Survey European Research Infrastructure, 2024)		Integer value between 0 and 10, where 0 represents "Cultural life is undermined" and 10 represents "Cultural life is enriched"

Question content	**What participants evaluate**	**(English translations of the) Response options**
Perceived effect of immigration on the Netherlands as a place to live (European Social Survey European Research Infrastructure, 2024)		Integer value between 0 and 10, where 0 represents "Worse place to live" and 10 represents "Better place to live"
Perceived tax burden		"...significantly too low", "...too low", "...just right", "...too high", "...significantly too high"
View on economic state intervention (question from the economic left-right scale of the EES; Schmitt et al., 2016)		Integer value between 0 and 10, where 0 represents "You are completely in favor of state intervention in the economy" and 10 represents "You are completely against state intervention in the economy"
View on economic redistribution (question from the economic left-right scale of the EES; Schmitt et al., 2016)		Integer value between 0 and 10, where 0 represents "You are completely in favor of redistributing wealth from the rich to the poor in the Netherlands" and 10 represents "You are completely against redistributing wealth from the rich to the poor in the Netherlands"

Question content	**What participants evaluate**	**(English translations of the) Response options**
View on taxation versus public services (question from the economic left-right scale of the EES; Schmitt et al., 2016)		Integer value between 0 and 10, where 0 represents "You are completely in favor of raising taxes to expand public services" and 10 represents "You are completely in favor of cutting public services to lower taxes"
Perceived frequency with which Dutch politicians speak to voters		"...too rarely", "...rarely", "...often", "...too often"
Perceived urgency of recruiting more soldiers		"Very urgent", "Moderately urgent", "Somewhat urgent", "Not urgent at all"

Question content	**What participants evaluate**	**(English translations of the) Response options**
Trust on Government (Grimmelikhuijsen, 2012), perceived transparency (Park & Blenkinsopp, 2011) and perceived responsiveness (Vigoda-Gadot & Yuval, 2003) of the government	"The government is responsive to public opinion", "Government institutions are honest", "Citizens' requests to the government are handled appropriately within a reasonable timeframe", "Government institutions are sincere", "The government shares sufficient information with the public about its performance", "Government institutions perform their tasks effectively", "The way the government works is made transparently known", "The government responds quickly to public requests", "The public can clearly see the government's progress", "Government institutions are competent", "The government efficiently reaches good solutions for general interests", "The government sincerely tries to help residents who need help", "Government institutions have the general interest in mind", "Government policy is implemented transparently"	"Strongly disagree", "Disagree", "Neither disagree nor agree", "Agree", "Strongly agree" (We adopted this unified response scheme for these 3 scales from Mansoor, 2021)

Question content	**What participants evaluate**	**(English translations of the) Response options**
Firmness of their personal political convictions		"Unshakable", "Firm", "Somewhat flexible", "Not at all firm"
Trust in institutions (European Social Survey European Research Infrastructure, 2024)	"the Dutch parliament", "the current Dutch government" [not part of the ESS, we added this item], "the legal system", "the police", "politicians", "political parties", "the European Parliament", "the United Nations", "Dutch newspapers" [not part of the ESS, we added this item]	Integer value between 0 and 10, where 0 represents "No trust at all" and 10 represents "Complete trust"; participants could also answer "Don't know"
Interest in news about foreign countries	"Belgium", "Germany", "United Kingdom", "Spain", "France", "United States of America", "Turkey", "Israel", "Ukraine", "Russia", "China", "India", "Indonesia", "Morocco", "South Africa"	"Not at all", "Little", "Somewhat", "A great deal"
Expected economic development of foreign countries	"Belgium", "Germany", "United Kingdom", "Spain", "France", "United States of America", "Turkey", "Israel", "Ukraine", "Russia", "China", "India", "Indonesia", "Morocco", "South Africa"	"Will become less prosperous", "Will be (about) equally prosperous", "Will become more prosperous", "Unsure"

Question content	**What participants evaluate**	**(English translations of the) Response options**
Whether they personally know a scientist		"No", "Yes, a scientist working for a company", "Yes, a scientist working for a university", "Yes, but they don't work for a company or university", "Yes, I am a scientist"

Question content	**What participants evaluate**	**(English translations of the) Response options**
Faith in Science (Farias et al., 2013; Rutjens et al., 2018)	"The only real kind of knowledge we can have is scientific knowledge", "In a demon-haunted world, science is a candle in the dark", "Scientists and science should get more respect in modern society", "All problems facing humans can be solved by science", "Science tells us everything there is to know about what reality consists of", "Science is the most valuable part of human culture", "Science gives us a better understanding of the universe than religion", "We believe too much in science and not enough in feelings and faith", "The scientific method is the only reliable way to obtain knowledge", "We can only rationally believe in what can be scientifically proven", "Science is the most efficient way to obtain truth"	"Strongly disagree", "Disagree", "Somewhat disagree", "Neither agree nor disagree", "Somewhat agree", "Agree", "Strongly agree"

Question content	**What participants evaluate**	**(English translations of the) Response options**
Agreement with individual science-related statements	"I know how the process of science works", "There is no knowledge that we should not seek", "Scientists almost never look at topics that could improve my life", "Scientists are completely disconnected from ordinary people", "Research is better placed in the hands of companies than universities", "Science can pose a danger to human existence", "I wish I could participate more actively in scientific research", "Rare important discoveries distract from the fact that most science is a waste of money", "Ordinary citizens should be able to vote on which topics research is conducted on"	"Strongly disagree", "Disagree", "Somewhat disagree", "Neither agree nor disagree", "Somewhat agree", "Agree", "Strongly agree"

Question content	**What participants evaluate**	**(English translations of the) Response options**
Perceived impact of sciences on their life	"Artificial Intelligence", "Archaeology", "Biology", "Chemistry", "Economics", "Pedagogy", "Engineering sciences", "Environmental science", "Linguistics", "Medicine", "Philosophy", "Physics", "Psychology"	"I don't know", "It (virtually) did not impact my life", "On aggregate, it worsened my life", "On aggregate, it had a neutral impact on my life", "On aggregate, it improved my life"
Whether a field is a "real science"	"Artificial Intelligence", "Archaeology", "Biology", "Chemistry", "Economics", "Pedagogy", "Engineering sciences", "Environmental science", "Linguistics", "Medicine", "Philosophy", "Physics", "Psychology"	"Yes", "No", "Don't know"

C.4 Appendix: Questions about news events

Click here to return to Chapter 6.

In the 3rd section of the questionnaire (on the content of newspapers), each block concluded with a question about whether certain news items related to a topic had taken place. This recurring question always had the same response options: "I am certain this happened", "I think this happened but I am not certain", "I don't think this happened but I am not certain", or "I am certain this did not happen".

Below are all news items, in order of occurrence, as well whether they are factual or fictitious.

News event	**Veracity**
Sport	
According to the CBS (Centraal Bureau voor de Statistiek), a record low number of Dutch adults reported an "affiliation to" or "liking of" a professional sports team in 2024.	fictitious
Femke Bol was caught using a performance-enhancing substance at the 400 m hurdle European championship in Rome.	fictitious
Max Verstappen became 4-time champion of the Formula 1.	factual
Challenger Gukesh Dommaraju became new chess world champion by defeating defending champion Ding Liren.	factual
PSV Eindhoven lead the Eredivisie table for the entire month of November.	factual
National coach Ronald Koeman came under fire from football legend Johan Cruyff after Holland's "disastrous" loss to Germany in the Nations League.	fictitious
Entertainment	
The movie "Joker: Folie à Deux" was a critical and commercial flop.	factual
The Concertgebouw orchestra spent most of November on a tour through the USA.	factual

News event	**Veracity**
Drake announced that he will sue fellow rapper Kendrick Lamar, because the latter accused him of being a pedophile.	fictitious as stated; Drake sued Lamar's label rather than Lamar himself
Arjen Lubach announced he will open his own TV channel "RTLu" in cooperation with RTL.	fictitious
Jimmy Kimmel was announced as host of the Academy Awards ("Oscars") in 2025.	fictitious
Breda-born Sylvie Meis won "critics' award" at the Milan Fashion Week.	fictitious
Netflix started cracking down on account sharing in the Netherlands.	factual
Domestic events and Domestic politics	
Mark Rutte became secretary general of NATO.	factual
In Ipsos's "Global Happiness 2024" report, the Netherlands ranked first (of 30).	factual
The Dutch government announced plans to severely reduce the number of foreign-language-taught Bachelor courses to reduce the number of international students.	factual
The Erasmusbrug was damaged after a collision with a yacht, requiring costly repairs through most of autumn.	fictitious
On the day of match between Ajax Amsterdam and Maccabi Tel Aviv, there were multiple attacks on Israeli soccer fans in Amsterdam.	factual
The Dutch government announced their support for immigration restrictions to ensure that there are at most 20 million people living in the Netherlands (by 2050).	factual
The RIVM (Centre for Infectious Disease Control) announced that 2024 was the lowest year on record for reported Rubella ("rodehond") cases in the Netherlands.	fictitious

News event	**Veracity**
Ten-thousands of students and academics gathered in Utrecht to protest the Dutch government's planned budget cuts to higher education.	intended factual, but item text contained an error: despite initial Utrecht plans, the protest ultimately took place in Den Haag
The Dutch government announced cuts to the sales tax to support middleclass families.	fictitious
International events and International politics	
German Chancellor Olaf Scholz lost a "vote of no confidence", paving the way for a federal election in 2025.	factual
President of South Korea Yoon Suk Yeol declared martial law—a decision that was reversed by the South Korean parliament within 6 hours.	factual
Following Xi Jinping's surprise retirement announcement, the Chinese Communist Party (CCP) announced that Vice President Han Zheng will succeed him in 2026.	fictitious
Sweden withdrew its veto, paving the way for Turkey's accession to NATO by the end of 2025.	fictitious
Former president Jimmy Carter died, aged 100.	factual
Romania's constitutional court annulled the second round of presidential elections in Romania.	factual
Negotiations about a national union of Australia and New Zealand collapsed after the first rounds of talks.	fictitious
Donald Trump won the 2024 election for US president by defeating Democratic candidate Kamala Harris.	factual
Israel performed multiple airstrikes on military targets in Syria, ostensibly to prevent Syria's new provisional government from acquiring the there-stored weapons.	factual

News event	**Veracity**
For the first time in South Africa's (democratic) history, the ANC party failed to win a majority of the vote in a general election.	factual
Science	
2024's Nobel prize in Physics was awarded to two researchers for their contributions to machine learning (a type of AI).	factual
OpenAI upgraded their application ChatGPT with GPT5.	fictitious (at the time)
In a successful testrun, a prototype quantum battery powered the Wanhua District of Taipei for three hours.	fictitious
In a testrun, a promising new drug, lenacapavir, prevented HIV infections for 6 months after injection.	factual
A new rotation-free method of creating artificial gravity was successfully tested on the ISS (International Space Station).	fictitious
As the first successful medical application of 3D-printed organs, a bladder was successfully transplanted to a pig.	fictitious
2024 was the hottest year on record, overtaking 2023.	factual
Scientists published, for the first time ever, a complete mapping of a brain. It was the brain of a fruit fly.	factual

Bibliography

Abbasi, A., Dobolyi, D., Lalor, J. P., Netemeyer, R. G., Smith, K., & Yang, Y. (2021). Constructing a Psychometric Testbed for Fair Natural Language Processing. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3748–3758. https://doi.org/10.18653/v1/2021.emnlp-main.304

Abele, A. E., Hauke, N., Peters, K., Louvet, E., Szymkow, A., & Duan, Y. (2016). Facets of the fundamental content dimensions: Agency with competence and assertiveness—communion with warmth and morality. *Frontiers in psychology*, *7*, 1810.

Abele, A. E., Uchronski, M., Suitner, C., & Wojciszke, B. (2008). Towards an operationalization of the fundamental dimensions of agency and communion: Trait content ratings in five countries considering valence and frequency of word occurrence. *European Journal of Social Psychology*, *38*(7), 1202–1217.

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology*, *93*(5), 751.

Abele, A. E., & Wojciszke, B. (2014). Communal and agentic content in social cognition: A dual perspective model. In *Advances in experimental social psychology* (pp. 195–255, Vol. 50). Elsevier.

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-muslim bias in large language models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306.

Adachi, T. (2013). Occupational gender stereotypes: Is the ratio of women to men a powerful determinant? *Psychological Reports*, *112*(2), 640–650.

Akour, M., & AL-Omari, H. (2013). Empirical Investigation of the Stability of IRT Item-Parameters Estimation. *International Online Journal of Educational Sciences*, 291–301. https://www.acarindex.com/dosyalar/makale/acarindex-1423904223.pdf

Akyürek, A. F., Kocyigit, M. Y., Paik, S., & Wijaya, D. T. (2022). Challenges in Measuring Bias via Open-Ended Language Generation. *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 76–76. https://doi.org/10.18653/v1/2022.gebnlp-1.9

Akyürek, A. F., Paik, S., Kocyigit, M., Akbiyik, S., Runyun, S. L., & Wijaya, D. (2022). On measuring social biases in prompt-based multi-task learning. *Findings of the Association for Computational Linguistics: NAACL 2022*, 551–564.

Amidei, J., Piwek, P., & Willis, A. (2020). Identifying Annotator Bias: A new IRT-based method for bias identification. *Proceedings of the 28th International Conference on Computational Linguistics*, 4787–4797. https://doi.org/10.18653/v1/2020.coling-main.421

Antoniak, M., & Mimno, D. (2021). Bad seeds: Evaluating lexical methods for bias measurement. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1889–1904.

Anunciacao, L. (2018). An overview of the history and methodological aspects of psychometrics: History and methodological aspects of psychometrics. *Journal for ReAttach Therapy and Developmental Diversities*, *1*(1), 44–58. https://www.ieb.usp.br/wp-content/uploads/sites/392/2018/08/History-and-Methodological-Aspects-of-Psychometrics-.pdf

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.

Aribandi, V., Tay, Y., & Metzler, D. (2021). How Reliable are Model Diagnostics? *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1778–1785. https://doi.org/10.18653/v1/2021.findings-acl.155

Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, *34*(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

Asgari, A., Wu, H., Naziri, A., Kolahdouzi, M., & Seyyed-Kalantari, L. (2026). Quantifying metric and model agreement in bias evaluation of large language models. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16868–16933.

Autoriteit Persoonsgegevens. (2025, October). *RAN special: AI-chatbots als stemhulp* (tech. rep.). Autoriteit Persoonsgegevens, Directie Coördinatie Algoritmes. Retrieved August 11, 2026, from https://www.autoriteitpersoonsgegevens.nl/documenten/ran-special-ai-chatbots-als-stemhulp

Bach, S., Sanh, V., Yong, Z. X., Webson, A., Raffel, C., Nayak, N. V., Sharma, A., Kim, T., Bari, M. S., Févry, T., et al. (2022). Promptsource: An integrated development environment and repository for natural language prompts. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 93–104.

Bachmann, D., van der Wal, O., Chvojka, E., Zuidema, W. H., van Maanen, L., & Schulz, K. (2024). Fl-irt-ing with psychometrics to improve nlp bias measurement: D. bachmann et al. *Minds and Machines*, *34*(4), 37.

Balayn, A., & Gürses, S. (2021). Beyond debiasing: Regulating ai and its inequalities. *EDRi Report*. https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf

Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science*, *7*(3), 136–141.

Barocas, S., Crawford, K., Shapiro, A., & Wallach, H. (2017). The problem with bias: Allocative versus representational harms in machine learning. *9th Annual conference of the special interest group for computing, information and society*.

Bartl, M., Nissim, M., & Gatt, A. (2020). Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. *arXiv preprint arXiv:2010.14534*.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bauer, P. C., Barberá, P., Ackermann, K., & Venetz, A. (2017). Is the left-right scale a valid measure of ideology? individual-level variation in associations with "left" and "right" and left-right self-placement. *Political Behavior*, *39*(3), 553–583.

Beeching, E., Fourrier, C., Habib, N., Han, S., Lambert, N., Rajani, N., Sanseviero, O., Tunstall, L., & Wolf, T. (2023). Open LLM leaderboard [Accessible on https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard].

Beggs, J. M., & Doolittle, D. C. (1993). Perceptions now and then of occupational sex typing: A replication of shinar's 1975 study. *Journal of Applied Social Psychology*, *23*(17), 1435–1453.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bhowmick, P. K., Basu, A., & Mitra, P. (2008). An Agreement Measure for Determining Inter-Annotator Reliability of Human Judgements on Affective Text. *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, 58–65. Retrieved October 4, 2023, from https://aclanthology.org/W08-1209

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., & van der Wal, O. (2023). Pythia: A suite for analyzing large language models across training and scaling. *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, *202*, 2397–2430. https://proceedings.mlr.press/v202/biderman23a.html

Binet, A. (1911). Nouvelles recherches sur la mesure du niveau intellectuel chez les enfants d'école. *L'Année psychologique*, *17*, 145–201. https://doi.org/10.3406/psy.1910.7275

Binet, A., & Simon, T. (1908). Le développement de l'intelligence chez les enfants. *L'Année psychologique*, *14*, 1–94. https://doi.org/10.3406/psy.1907.3737

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, *32*(4), 401–416.

Bitton, Y., Bitton, E., & Nisan, S. (2025). Detecting stylistic fingerprints of large language models. *arXiv preprint arXiv:2503.01659*.

Bleidorn, W., Lenhausen, M. R., Schwaba, T., Gebauer, J. E., & Hopwood, C. J. (2023). Secularization trends obscure developmental changes in religiosity. *Social Psychological and Personality Science*, *14*(2), 249–258.

Bleidorn, W., Schwaba, T., Zheng, A., Hopwood, C. J., Sosa, S. S., Roberts, B. W., & Briley, D. (2022). Personality stability and change: A meta-analysis of longitudinal studies. *Psychological bulletin*, *148*(7-8), 588.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., & Wallach, H. (2021). Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1004–1015. https://doi.org/10.18653/v1/2021.acl-long.81

Boag, R. J., Innes, R. J., Stevenson, N., Bahg, G., Busemeyer, J. R., Cox, G. E., Donkin, C., Frank, M. J., Hawkins, G. E., Heathcote, A., et al. (2025). An expert guide to planning experimental tasks for evidence-accumulation modeling. *Advances in Methods and Practices in Psychological Science*, *8*(2), 25152459251336127.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 4356–4364.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). On the Opportunities and Risks of Foundation Models. https://doi.org/10.48550/arXiv.2108.07258

Bommasani, R., & Liang, P. (2022). Trustworthy Social Bias Measurement. https://doi.org/10.48550/arXiv.2212.11672

Bonifay, W. (2019). *Multidimensional Item Response Theory*. Sage Publications.

Bordia, S., & Bowman, S. R. (2019). Identifying and Reducing Gender Bias in Word-Level Language Models. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 7–15. https://doi.org/10.18653/v1/N19-3002

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity [Place: US Publisher: American Psychological Association]. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Brewer, N., Zoanetti, J., & Young, R. L. (2017). The influence of media suggestions about links between criminality and autism spectrum disorder. *Autism*, *21*(1), 117–121.

Broockman, D., & Kalla, J. (2022). The manifold effects of partisan media on viewers' beliefs and attitudes: A field experiment with fox news viewers. *OSF preprints*, *1*, 1–42.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. Retrieved July 5, 2023, from https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Brysbaert, M., & Nicolas, S. (2024). Two persistent myths about binet and the beginnings of intelligence tests in psychology textbooks. *Collabra: Psychology*, *10*(1), 117600.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Bürki, A. (2025). Assessing effect sizes, variability, and power in the web-based study of language production. *Language, Cognition and Neuroscience*, *40*(10), 1291–1307.

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Byrd, M., & Srivastava, S. (2022). Predicting difficulty and discrimination of natural language questions. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 119–130. https://aclanthology.org/2022.acl-short.15/

Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item Response Theory. *Annual Review of Statistics and Its Application*, *3*, 297–321. https://doi.org/10.1146/annurev-statistics-041715-033702

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review*, *16*(4), 330–350.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81–105.

Cao, Y., Pruksachatkun, Y., Chang, K.-W., Gupta, R., Kumar, V., Dhamala, J., & Galstyan, A. (2022). On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 561–570. https://doi.org/10.18653/v1/2022.acl-short.62

Casper, C., Rothermund, K., & Wentura, D. (2010). Automatic stereotype activation is context dependent. *Social Psychology*.

Cerina, R. (2025). Possum: A protocol for surveying social-media users with multimodal llms. *arXiv preprint arXiv:2503.05529*.

Chalmers, R. P. (2012). mirt: A multidimensional Item Response Theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29. https://doi.org/10.18637/jss.v048.i06

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Aden-Buie, G., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2026). *Shiny: Web application framework for r* [R package version 1.13.0.9000]. https://github.com/rstudio/shiny

Chen, Y., Raghuram, V. C., Mattern, J., Mihalcea, R., & Jin, Z. (2025). Causally testing gender bias in llms: A case study on occupational bias. *Findings of the Association for Computational Linguistics: NAACL 2025*, 4999–5019.

Cheng, L., Varshney, K. R., & Liu, H. (2021). Socially Responsible AI Algorithms: Issues, Purposes, and Challenges. *Journal of Artificial Intelligence Research*, *71*, 1137–1181. https://doi.org/10.1613/jair.1.12814

Chuey, A., Boyce, V., Cao, A., & Frank, M. C. (2024). Conducting developmental research online vs. in-person: A meta-analysis. *Open Mind*, *8*, 795–808.

Ciora, C., Iren, N., & Alikhani, M. (2021). Examining covert gender bias: A case study in turkish and english machine translation models. *Proceedings of the 14th International Conference on Natural Language Generation*, 55–63.

Coppock, A., Ekins, E., & Kirby, D. (2018). The long-lasting effects of newspaper op-eds on public opinion. *Quarterly Journal of Political Science*, *13*(1), 59–87. https://doi.org/10.1561/100.00016112

Corneille, O., & Gawronski, B. (2024). Self-reports are better measurement instruments than implicit measures. *Nature Reviews Psychology*, *3*(12), 835–846.

Costa Jr, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of personality and social psychology*, *81*(2), 322.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. ERIC.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests [Place: US Publisher: American Psychological Association]. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., ... Sculley, D. (2022). Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research*, *23*(226), 1–61. http://jmlr.org/papers/v23/20-1335.html

Danesi, M. (2014). *Dictionary of media and communications*. Routledge.

Dassonneville, R. (2021). Change and continuity in the ideological gender gap a longitudinal analysis of left-right self-placement in oecd countries. *European Journal of Political Research*, *60*(1), 225–238.

Dastin, J. (2018, October 11). *Amazon scraps secret ai recruiting tool that showed bias against women* [Archived at https://perma.cc/B8MV-THAD]. Reuters.

De Ayala, R. J. (2013). *The Theory and Practice of Item Response Theory*. Guilford Publications.

De Cao, N., Schmid, L., Hupkes, D., & Titov, I. (2022). Sparse Interventions in Language Models with Differentiable Masking. *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 16–27. Retrieved July 5, 2023, from https://aclanthology.org/2022.blackboxnlp-1.2

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "small world of words" english word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987–1006.

De Vries, C. E. (2018). The cosmopolitan-parochial divide: Changing patterns of party and electoral competition in the netherlands and beyond. *Journal of European Public Policy*, *25*(11), 1541–1565.

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., & Kalai, A. T. (2019). Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 120–128. https://doi.org/10.1145/3287560.3287572

Debelak, R., Strobl, C., & Zeigenfuse, M. D. (2022). *An Introduction to the Rasch Model with Examples in R*. CRC Press.

Dehghani, M., Tay, Y., Gritsenko, A. A., Zhao, Z., Houlsby, N., Diaz, F., Metzler, D., & Vinyals, O. (2021). The benchmark lottery. *arXiv preprint arXiv:2107.07002*.

Delobelle, P., Tokpo, E., Calders, T., & Berendt, B. (2022). Measuring Fairness with Biased Rulers: A Comparative Study on Bias Metrics for Pre-trained Language Models. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1693–1706. https://aclanthology.org/2022.naacl-main.122

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, *13*(4), 354–378. https://doi.org/10.1080/15305058.2013.799067

Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J., & Chang, K.-W. (2021). Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1968–1994. https://doi.org/10.18653/v1/2021.emnlp-main.150

Dev, S., Sheng, E., Zhao, J., Amstutz, A., Sun, J., Hou, Y., Sanseverino, M., Kim, J., Nishi, A., Peng, N., & Chang, K.-W. (2022). On Measures of Biases and Harms in NLP. *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, 246–267. Retrieved July 5, 2023, from https://aclanthology.org/2022.findings-aacl.24

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., & Gupta, R. (2021). BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 862–872. https://doi.org/10.1145/3442188.3445924

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., & Williams, A. (2020). Multi-Dimensional Gender Bias Classification. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 314–331. https://doi.org/10.18653/v1/2020.emnlp-main.23

Djourelova, M. (2023). Persuasion through slanted language: Evidence from the media coverage of immigration. *American economic review*, *113*(3), 800–835.

Döll, M., Döhring, M., & Müller, A. (2024). Evaluating gender bias in large language models. *arXiv preprint arXiv:2411.09826*.

Dovidio, J. F., Evans, N., & Tyler, R. B. (1986). Racial stereotypes: The contents of their cognitive representations. *Journal of experimental social psychology*, *22*(1), 22–37.

Dražanová, L., Gonnot, J., Heidland, T., & Krüger, F. (2024). Which individual-level factors explain public attitudes toward immigration? a meta-analysis. *Journal of ethnic and migration studies*, *50*(2), 317–340.

Druckman, J. N., & Parkin, M. (2005). The impact of media bias: How editorial slant affects voters. *The Journal of Politics*, *67*(4), 1030–1049.

Du, Y., Fang, Q., & Nguyen, D. (2021). Assessing the Reliability of Word Embedding Gender Bias Measures. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10012–10034. https://doi.org/10.18653/v1/2021.emnlp-main.785

Durmus, E., Nguyen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., et al. (2023). Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changed: A cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018 [place: US publisher: American Psychological Association]. *American Psychologist*, *75*, 301–315. https://doi.org/10.1037/amp0000494

Elbouanani, A., Dufraisse, E., & Popescu, A. (2025). Analyzing political bias in llms via target-oriented sentiment classification. *Findings of the Association for Computational Linguistics: ACL 2025*, 15476–15505.

Ellemers, N. (2018). Gender stereotypes. *Annual Review of Psychology*, *69*(1), 275–298.

Elvestad, E., & Blekesaune, A. (2008). Newspaper readers in europe: A multilevel study of individual and national differences. *European Journal of Communication*, *23*(4), 425–447.

Entman, R. M. (2007). Framing bias: Media in the distribution of power. *Journal of communication*, *57*(1), 163–173.

Ethayarajh, K. (2020). Is Your Classifier Actually Biased? Measuring Fairness under Uncertainty with Bernstein Bounds. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2914–2919. https://doi.org/10.18653/v1/2020.acl-main.262

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). Understanding Undesirable Word Embedding Associations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1696–1705. https://doi.org/10.18653/v1/P19-1166

European Social Survey European Research Infrastructure. (2024). Ess round 11 (2023/2024) [Integrated dataset, edition 2.0]. https://www.europeansocialsurvey.org/data-portal

Fang, Q., Nguyen, D., & Oberski, D. L. (2022). Evaluating the construct validity of text embeddings with application to survey questions. *EPJ Data Science*, *11*(1), 39. https://doi.org/10.1140/epjds/s13688-022-00353-7

Fang, Q., Oberski, D. L., & Nguyen, D. (2024). PATCH – psychometrics-assisted benchmarking of large language models: A case study of mathematics proficiency. https://arxiv.org/abs/2404.01799

Farias, M., Newheiser, A.-K., Kahane, G., & de Toledo, Z. (2013). Scientific faith: Belief in science increases in the face of stress and existential anxiety. *Journal of experimental social psychology*, *49*(6), 1210–1213.

Farkas, A., & Németh, R. (2022). How to measure gender bias in machine translation: Real-world oriented machine translators, multiple reference points. *Social Sciences & Humanities Open*, *5*(1), 100239.

Field, A., Blodgett, S. L., Waseem, Z., & Tsvetkov, Y. (2021). A Survey of Race, Racism, and Anti-Racism in NLP. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1905–1925. https://doi.org/10.18653/v1/2021.acl-long.149

Field, A., & Tsvetkov, Y. (2020). Unsupervised Discovery of Implicit Gender Bias. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 596–608. https://doi.org/10.18653/v1/2020.emnlp-main.44

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in methods and practices in psychological science*, *3*(4), 456–465.

Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, *75*(1), 241–268.

Fortunati, L., Deuze, M., & De Luca, F. (2014). The new about news: How print, online, free, and mobile coconstruct new audiences in italy, france, spain, the uk, and germany. *Journal of Computer-Mediated Communication*, *19*(2), 121–140.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the international AAAI conference on web and social media*, *12*(1).

Fox, J.-P. (2010). *Introduction to Bayesian Response Modeling*. Springer.

Frank, M. C. (2023). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, *2*(8), 451–452.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (Im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, *64*(4), 136–143. https://doi.org/10.1145/3433949

Frigg, R., & Hartmann, S. (2006). Models in science [First published Feb 27, 2006; substantive revision Jun 25, 2012]. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/models-science/

Furr, R. M. (2021). *Psychometrics: An Introduction* (Fourth). SAGE Publications.

Gaertner, S. L., & McLaughlin, J. P. (1983). Racial stereotypes: Associations and ascriptions of positive and negative characteristics. *Social Psychology Quarterly*, 23–30.

Galobardes, B., Shaw, M., Lawlor, D. A., Lynch, J. W., & Davey Smith, G. (2006). Indicators of socioeconomic position (part 1). *Journal of epidemiology and community health*, *60*(1), 7–12.

Gao, C., Shinkareva, S. V., & Desai, R. H. (2023). Scope: The south carolina psycholinguistic metabase. *Behavior Research Methods*, *55*(6), 2853–2884.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. *Handbook of Research Methods in Social and Personality Psychology*, *2*, 283–310.

Gawronski, B., & Hahn, A. (2018). Implicit measures: Procedures, use, and interpretation. In *Measurement in social psychology* (pp. 29–55). Routledge.

Gelb, S. A. (1986). Henry h. goddard and the immigrants, 1910-1917: The studies and their social context. *Journal of the History of the Behavioral Sciences*, *22*(4), 324–332.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models.

Gerber, A. S., Karlan, D., & Bergan, D. (2009). Does the media matter? a field experiment measuring the effect of newspapers on voting behavior and political opinions. *American Economic Journal: Applied Economics*, *1*(2), 35–52. https://doi.org/10.1257/app.1.2.35

Gilholm, P., Mengersen, K., & Thompson, H. (2021). Bayesian hierarchical multidimensional item response modeling of small sample, sparse data for personalized developmental surveillance. *Educational and Psychological Measurement*, *81*(5), 936–956. https://doi.org/10.1177/0013164420987582

Goddard, H. H. (1917). Mental tests and the immigrant. *Journal of Delinquency*, *2*(5), 243–277.

Golchin, S., & Surdeanu, M. (2023). Time travel in llms: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*.

Goldfarb-Tarrant, S., Marchant, R., Muñoz Sánchez, R., Pandya, M., & Lopez, A. (2021). Intrinsic Bias Metrics Do Not Correlate with Application Bias. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1926–1940. https://doi.org/10.18653/v1/2021.acl-long.150

Goldfarb-Tarrant, S., Ungless, E. L., Balkir, E., & Blodgett, S. L. (2023). This prompt is measuring <MASK>: Evaluating bias evaluation in language models. *Findings of the Association for Computational Linguistics: ACL 2023*, 2209–2225.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 609–614. https://doi.org/10.18653/v1/N19-1061

Gonen, H., Ravfogel, S., & Goldberg, Y. (2022). Analyzing Gender Representation in Multilingual Models. *Proceedings of the 7th Workshop on Representation Learning for NLP*, 67–77. https://doi.org/10.18653/v1/2022.repl4nlp-1.8

Gottfried, J., Bishop, W., Anderson, M., Faverio, M., Park, E., & McClain, C. (2026, June 17). *Americans and AI 2026: Chatbots, smart devices and views on impact* (Research Report). Pew Research Center. https://web.archive.org/web/20260725193208/https://www.pewresearch.org/internet/2026/06/17/americans-and-ai-2026-chatbots-smart-devices-and-views-on-impact/

Greenberg, S. (2024). Always conduct the "simplest valid analysis" [Accessed: 2025-11-24]. https://web.archive.org/web/20251124150317/https://replications.clearerthinking.org/simplest-valid-analysis/

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity [Place: US Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *97*(1), 17–41. https://doi.org/10.1037/a0015575

Grimm, P. (2010). Social desirability bias. *Wiley international encyclopedia of marketing*.

Grimmelikhuijsen, S. (2012). Linking transparency, knowledge and citizen trust in government: An experiment. *International Review of Administrative Sciences*, *78*(1), 50–73.

Grün, B., & Bürkner, P.-C. (2020). Bridgesampling: An r package for estimating normalizing constants. *Journal of Statistical Software*, *92*(10), 1–29. https://doi.org/10.18637/jss.v092.i10

Haines, E. L., Deaux, K., & Lofaro, N. (2016). The times they are a-changing… or are they not? a comparison of gender stereotypes, 1983–2014. *Psychology of Women Quarterly*, *40*(3), 353–363.

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.

Hanitzsch, T., & Vos, T. P. (2018). Journalism beyond democracy: A new look into journalistic roles in political and everyday life. *Journalism*, *19*(2), 146–164.

Hanusch, F. (2019). Journalistic roles and everyday life: An empirical account of lifestyle journalists' professional views. *Journalism Studies*, *20*(2), 193–211.

Harrington, D. (2009). *Confirmatory Factor Analysis*. Oxford University Press, USA.

Harteveld, E. (2021). Fragmented foes: Affective polarization in the multiparty context of the netherlands. *Electoral Studies*, *71*, 102332.

Hayes, B. C., & Tariq, V. N. (2000). Gender differences in scientific knowledge and attitudes toward science: A comparative study of four anglo-american nations. *Public Understanding of Science*, *9*(4), 433–447. https://doi.org/10.1088/0963-6625/9/4/306

Heilman, M. E., Caleo, S., & Manzi, F. (2024). Women at work: Pathways from gender stereotypes to gender bias and discrimination. *Annual Review of Organizational Psychology and Organizational Behavior*, *11*(1), 165–192.

Hladká, A., & Martinková, P. (2020). difNLR: Generalized logistic regression models for DIF and DDF detection. *R Journal*, *12*(1). https://journal.r-project.org/archive/2020/RJ-2020-014/

Hogenboom, S. A., Schulz, K., & van Maanen, L. (2024). Implicit association tests: Stimuli validation from participant responses. *British Journal of Social Psychology*, *63*(2), 975–1002.

Höhne, J. K., Krebs, D., & Kühnel, S.-M. (2021). Measurement properties of completely and end labeled unipolar and bipolar scales in likert-type questions on income (in) equality. *Social science research*, *97*, 102544.

Hooker, S., Moorosi, N., Clark, G., Bengio, S., & Denton, E. (2020). Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*.

Huang, J., Jiao, W., Lam, M. H., Li, E. J., Wang, W., & Lyu, M. (2024). On the reliability of psychological scales on large language models. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 6152–6173.

Huang, T., & Rothermund, K. (2022). Automatic activation of age stereotypes. *Social Psychology*.

Huang, T., & Rothermund, K. (2023). Endorsement and embodiment of cautiousness-related age stereotypes. *Frontiers in psychology*, *14*, 1091763.

Hussey, I., Alsalti, T., Bosco, F., Elson, M., & Arslan, R. (2025). An aberrant abundance of cronbach's alpha values at. 70. *Advances in Methods and Practices in Psychological Science*, *8*(1), 25152459241287123.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social biases in NLP models as barriers for persons with disabilities. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5491–5501. https://doi.org/10.18653/v1/2020.acl-main.487

ISSP Research Group. (1995). International social survey programme: Environment i – issp 1993. https://doi.org/10.4232/1.2450

Jacobs, A. Z., & Wallach, H. (2021). Measurement and Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385. https://doi.org/10.1145/3442188.3445901

JASP Team. (2025). JASP (Version 0.95.4)[Computer software]. https://jasp-stats.org/

Jiao, M., & Luo, Z. (2021). Gender Bias Hidden Behind Chinese Word Embeddings: The Case of Chinese Adjectives. *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, 8–15. https://doi.org/10.18653/v1/2021.gebnlp-1.2

Jim, J. R., Talukder, M. A. R., Malakar, P., Kabir, M. M., Nur, K., & Mridha, M. F. (2024). Recent advancements and challenges of nlp-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, *6*, 100059.

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed.). Retrieved April 26, 2023, from https://web.stanford.edu/~jurafsky/slp3/

Kaiser, C., Kaiser, J., Manewitsch, V., Rau, L., & Schallner, R. (2025). Simulating human opinions with large language models: Opportunities and challenges for personalized survey data modeling. *Adjunct proceedings of the 33rd ACM conference on user modeling, adaptation and personalization*, 82–86.

Kalogeropoulos, A., Toff, B., & Fletcher, R. (2024). The watchdog press in the doghouse: A comparative study of attitudes about accountability journalism, trust in news, and news avoidance. *The International Journal of Press/Politics*, *29*(2), 485–506.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of educational measurement*, *50*(1), 1–73.

Karylowski, J. J., Motes, M. A., Wallace, H. M., Harckom, H. A., Hewlett, E. M., Maclean, S. L., Vaswani, C., et al. (2001). Spontaneous gender-stereotypical categorization of trait labels and job labels. *Current Research in Social Psychology*, *6*(6), 77–90.

Kawakami, K., & Dovidio, J. F. (2001). The reliability of implicit stereotyping. *Personality and Social Psychology Bulletin*, *27*(2), 212–225.

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. *Proceedings of the 33rd annual acm conference on human factors in computing systems*, 3819–3828.

Keita, S., Renault, T., & Valette, J. (2024). The usual suspects: Offender origin, media reporting and natives' attitudes towards immigration. *The Economic Journal*, *134*(657), 322–362.

Khanna, S., & Li, X. (2025). Invisible languages of the llm universe. *arXiv preprint arXiv:2510.11557*. https://doi.org/10.48550/arXiv.2510.11557

Kidd, C., & Birhane, A. (2023). How AI can distort human beliefs. *Science*, *380*(6651), 1222–1223.

Kidder, C. K., White, K. R., Hinojos, M. R., Sandoval, M., & Crites Jr, S. L. (2018). Sequential stereotype priming: A meta-analysis. *Personality and Social Psychology Review*, *22*(3), 199–227.

Kiritchenko, S., Nejadgholi, I., & Fraser, K. C. (2021). Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective. *Journal of Artificial Intelligence Research*, *71*, 431–478. https://doi.org/10.1613/jair.1.12590

Kline, P. (2014). *An easy guide to factor analysis*. Routledge.

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, *57*(2), 1813–1824. https://doi.org/10.1016/j.compedu.2011.02.003

König, C., Spoden, C., & Frey, A. (2020). An optimized Bayesian hierarchical two-parameter logistic model for small-sample item calibration. *Applied Psychological Measurement*, *44*(4), 311–326. https://doi.org/10.1177/0146621619893786

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & quantity*, *47*(4), 2025–2047.

Lalor, J. P., Wu, H., & Yu, H. (2016). Building an evaluation scale using item response theory. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, *2016*, 648.

Lalor, J. P., & Yu, H. (2020). Dynamic data selection for curriculum learning via ability estimation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, *2020*, 545. https://aclanthology.org/2020.findings-emnlp.48/

Lam, C. S. P., Arnott, C., Beale, A. L., Chandramouli, C., Hilfiker-Kleiner, D., Kaye, D. M., Ky, B., Santema, B. T., Sliwa, K., & Voors, A. A. (2019). Sex differences in heart failure. *European Heart Journal*, *40*(47), 3859–3868c. https://doi.org/10.1093/eurheartj/ehz835

Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S., & Huang, J. X. (2023). A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. *arXiv preprint arXiv:2305.18486*.

Lee, S., Peng, T.-Q., Goldberg, M. H., Rosenthal, S. A., Kotcher, J. E., Maibach, E. W., & Leiserowitz, A. (2024). Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *PLoS Climate*, *3*(8), e0000429.

Leidinger, A., Van Rooij, R., & Shutova, E. (2023). The language of prompting: What linguistic properties make a prompt successful? *Findings of the association for computational linguistics: EMNLP 2023*, 9210–9232.

Levy, S., Lazar, K., & Stanovsky, G. (2021). Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation

[tex.ids= levy2021CollectingLargeScaleGendera]. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2470–2480. https://doi.org/10.18653/v1/2021.findings-emnlp.211

Limisiewicz, T., & Mareček, D. (2022). Don't Forget About Pronouns: Removing Gender Bias in Language Models Without Losing Factual Gender Information. *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 17–29. https://doi.org/10.18653/v1/2022.gebnlp-1.3

Liscum, M., & Garcia, M. L. (2022). You can't keep a bad idea down: Dark history, death, and potential rebirth of eugenics. *The Anatomical Record*, *305*(4), 902–937.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, *55*(9), 195:1–195:35. https://doi.org/10.1145/3560815

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., & Roberts, A. (2023). The Flan collection: Designing data and methods for effective instruction tuning. *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, *202*, 22631–22648. https://proceedings.mlr.press/v202/longpre23a.html

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Routledge.

Lühiste, K. (2006). Explaining trust in political institutions: Some illustrations from the baltic states. *Communist and post-communist studies*, *39*(4), 475–496.

Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Analytic posteriors for pearson's correlation coefficient. *Journal of Mathematical Psychology*, *84*, 84–96. https://doi.org/10.1016/j.jmp.2018.03.006

Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized Adaptive and Multistage Testing with R: Using Packages catR and mstR*. Springer.

Malcolm X. (1965). *Malcolm X speaks: Selected speeches and statements* (G. Breitman, Ed.). Grove Press.

Malik, V., Dev, S., Nishi, A., Peng, N., & Chang, K.-W. (2022). Socially Aware Bias Measurements for Hindi Language Representations. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1041–1052. https://doi.org/10.18653/v1/2022.naacl-main.76

Mansoor, M. (2021). Citizens' trust in government as a function of good governance and government agency's provision of quality information on social media during covid-19. *Government information quarterly*, *38*(4), 101597.

Marlar, J., & Ritter, Z. (2026). Gallup begins research on simulated responses [Gallup Announcement]. Retrieved May 19, 2026, from https://web.archi

ve.org/web/20260519110236/https://news.gallup.com/opinion/methodology/709373/gallup-begins-research-synthetic-responses.aspx

Martinková, P., & Hladká, A. (2023). *Computational Aspects of Psychometric Methods: With R*. CRC Press.

Mathet, Y., Widlöcher, A., & Métivier, J.-P. (2015). The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, *41*(3), 437–479. https://doi.org/10.1162/COLI_a_00227

Mattern, J., Jin, Z., Sachan, M., Mihalcea, R., & Schölkopf, B. (2022). Understanding stereotypes in language models: Towards robust measurement and zero-shot debiasing. *ArXiv*, 2212–10678.

Maxwell, R. (2019). Cosmopolitan immigration attitudes in large european cities: Contextual or compositional effects? *American Political Science Review*, *113*(2), 456–474.

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). On Measuring Social Biases in Sentence Encoders. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 622–628. https://doi.org/10.18653/v1/N19-1063

Meade, N., Poole-Dayan, E., & Reddy, S. (2022). An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1878–1898. https://doi.org/10.18653/v1/2022.acl-long.132

Montuschi, P., Gatteschi, V., Lamberti, F., Sanna, A., & Demartini, C. (2013). Job recruitment and job seeking processes: How technology can help. *It professional*, *16*(5), 41–49.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to cousineau (2005). *Tutorials in quantitative methods for psychology*, *4*(2), 61–64.

Morey, R. D., & Rouder, J. N. (2023). *Bayesfactor: Computation of bayes factors for common designs* [R package version 0.9.12-4.6 (or later)]. https://CRAN.R-project.org/package=BayesFactor

Müller, F., & Rothermund, K. (2014). What does it take to activate stereotypes? simple primes don't seem enough. *Social Psychology*.

Munoz Sastre, M. T., Fouquereau, E., Igier, V., Salvatore, N., & Mullet, E. (2000). Perception of occupational gender typing: A replication on european samples of shinar's (1975) and beggs and doolittle's (1993) studies 1. *Journal of Applied Social Psychology*, *30*(2), 430–441.

Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th In-*

ternational Joint Conference on Natural Language Processing (Volume 1: Long Papers), 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1953–1967. https://doi.org/10.18653/v1/2020.emnlp-main.154

Nationaal Media Onderzoek. (2025, October). *NPMM 2025-II: Gemiddeld bereik editie (print + digitale replica) dagbladen* (NMO Print & Brands Monitor 2025-II; readership figures based on fieldwork covering Q2 2024 through Q1 2025). https://www.ndpnieuwsmedia.nl/wp-content/uploads/2025/10/Editiebereik-landelijke-en-regionale-dagbladen-NPMM-2025-II.pdf

Névéol, A., Dupont, Y., Bezançon, J., & Fort, K. (2022). French CrowS-Pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8521–8531. https://doi.org/10.18653/v1/2022.acl-long.583

Newton, K., & Brynin, M. (2001). The national press and party voting in the uk. *Political Studies*, *49*(2), 265–285. https://doi.org/10.1111/1467-9248.00313

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, *18*, 301–319. https://doi.org/10.1037/a0032969

Nguyen, D., & Ploeger, E. (2025). We need to measure data diversity in NLP —better and broader. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 8823–8832.

Nissim, M., van Noord, R., & van der Goot, R. (2020). Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, *46*(2), 487–497.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... Yarkoni, T. (2015). Promoting an open research culture [Publisher: American Association for the Advancement of Science]. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606.

Office for National Statistics. (2023). Employment by detailed occupation and sex (annual population survey, uk, 2023). https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes

/adhocs/2316employmentbydetailedoccupationandindustrybysexagegroupandcountry2020to2023

Orgad, H., & Belinkov, Y. (2022). Choose your lenses: Flaws in gender bias evaluation. *GeBNLP 2022*, 151.

Paek, I., & Cole, K. (2019). *Using R for Item Response Theory Model Applications*. Routledge.

Park, H., & Blenkinsopp, J. (2011). The roles of transparency and trust in the relationship between corruption and citizen satisfaction. *International Review of Administrative Sciences*, *77*(2), 254–274.

Parmar, M., Mishra, S., Geva, M., & Baral, C. (2023). Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions [tex.ids= parmar2023DonBlameAnnotatora arXiv: 2205.00415 [cs]]. Retrieved May 31, 2023, from http://arxiv.org/abs/2205.00415

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., & Bowman, S. R. (2021). BBQ: A hand-built bias benchmark for question answering. https://arxiv.org/abs/2110.08193v2

Peirce, J. W. (2007). Psychopy—psychophysics software in python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13.

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, *19*(5), 808–826.

Petty, R. E., Fazio, R. H., & Brinol, P. (2008). The new implicit measures: An overview. *Attitudes*, 23–38.

Pew Research Center. (2016, March). *The gender gap in religion around the world: Women are generally more religious than men, particularly among christians*. Pew Research Center. Washington, DC. Retrieved August 15, 2026, from https://web.archive.org/web/20260815124050/https://www.pewresearch.org/religion/2016/03/22/the-gender-gap-in-religion-around-the-world/

Pietraszkiewicz, A., Formanowicz, M., Gustafsson Sendén, M., Boyd, R. L., Sikström, S., & Sczesny, S. (2019). The big two dictionaries: Capturing agency and communion in natural language. *European Journal of Social Psychology*, *49*(5), 871–887.

Polo, F. M., Weber, L., Choshen, L., Sun, Y., Xu, G., & Yurochkin, M. (2024). tinyBenchmarks: Evaluating LLMs with fewer examples. https://arxiv.org/abs/2402.14992v1

Povinelli, D. (2012). *World without weight: Perspectives on an alien mind*. Oxford University Press.

Prates, M. O., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: A case study with google translate. *Neural Computing and Applications*, *32*(10), 6363–6381.

Prystawski, B., Grant, E., Nematzadeh, A., Lee, S. W. S., Stevenson, S., & Xu, Y. (2022). The Emergence of Gender Associations in Child Language Development. *Cognitive Science*, *46*(6), e13146. https://doi.org/10.1111/cogs.13146

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Riccardi, N., Yang, X., & Desai, R. H. (2024). The two word test as a semantic benchmark for large language models. *Scientific reports*, *14*(1), 21593.

Rijmen, F. (2011). Hierarchical factor Item Response Theory models for PIRLS: Capturing clustering effects at multiple levels. *IERI Monograph Series: Issues and Methodologies in Large-scale Assessments*, *4*, 59–74. https://ierinstitute.org/fileadmin/Documents/IERI_Monograph/Volume_4/IERI_Monograph_Volume_04_Chapter_3.pdf

Ringel, S. (2023). Digitizing the paper of record: Archiving digital newspapers at the new york times. *Journalism*, *24*(2), 245–261.

Rodriguez, P., Barrow, J., Hoyle, A. M., Lalor, J. P., Jia, R., & Boyd-Graber, J. (2021). Evaluation examples are not equally informative: How should that change NLP leaderboards? *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4486–4503. https://aclanthology.org/2021.acl-long.346

Rodriguez, P., Htut, P. M., Lalor, J. P., & Sedoc, J. (2022). Clustering examples in multi-dataset benchmarks with Item Response Theory. *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, 100–112. https://aclanthology.org/2022.insights-1.14/

Rosenbluth, T. (2026, July 22). *Chatgpt led to a man's near-fatal health crisis, lawsuit claims* [A shortened republication of this article is archived at https://web.archive.org/web/20260811080320/https://gvwire.com/2026/07/22/chatgpt-led-to-a-mans-near-fatal-health-crisis-lawsuit-claims/]. The New York Times. Retrieved August 11, 2026, from https://www.nytimes.com/2026/07/22/well/openai-chatgpt-health-lawsuit.html

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206–215.

Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). Gender Bias in Coreference Resolution. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 8–14. https://doi.org/10.18653/v1/N18-2002

Ruhl, C. (2023). *Demand characteristics in psychology*. Simply Psychology. https://web.archive.org/web/20260813140134/https://www.simplypsychology.org/demand-characteristics.html

Rutjens, B. T., Sutton, R. M., & van der Lee, R. (2018). Not all skepticism is equal: Exploring the ideological antecedents of science acceptance and rejection. *Personality and Social Psychology Bulletin*, *44*(3), 384–405.

Sahin, A., & Anil, D. (2017). The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory [Accepted: 2019-12-17T06:52:32Z Publisher: Edam]. https://doi.org/10.12738/estp.2017.1.0270

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*, 1–97. https://doi.org/10.1007/BF03372160

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Le Scao, T., Raja, A., et al. (2022). Multitask prompted training enables zero-shot task generalization. *ICLR 2022-Tenth International Conference on Learning Representations*.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The Risk of Racial Bias in Hate Speech Detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678. https://doi.org/10.18653/v1/P19-1163

Sauter, M., Stefani, M., & Mack, W. (2022). Equal quality for online and lab data: A direct comparison from two dual-task paradigms. *Open Psychology*, *4*(1), 47–59.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Schick, T., Udupa, S., & Schütze, H. (2021). Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, *9*, 1408–1424.

Schimmack, U. (2021). The implicit association test: A method in search of a construct. *Perspectives on Psychological Science*, *16*(2), 396–414.

Schmitt, H., Hobolt, S. B., Popa, S. A., Teperoglou, E., et al. (2016). European parliament election study 2014, voter study, first post-election survey. *GESIS Data Archive, Cologne. ZA5160 Data file Version*, *4*(0).

Schröder, A., Gemballa, T., Ruppin, S., & Wartenburger, I. (2012). German norms for semantic typicality, age of acquisition, and concept familiarity. *Behavior Research Methods*, *44*(2), 380–394.

Schulz, K., Lanzing, M., & Brenner, G. M. (2025). Moving beyond harm. a critical review of how nlp research approaches discrimination. *arXiv preprint arXiv:2508.04504*.

Schwegler, R. A. (1914). *A teachers' manual for the use of the binet-simon scale of intelligence* [Published as a University of Kansas, School of Education publication]. Kansas State Printing Office. Retrieved July 23, 2026, from https://web.archive.org/web/20260724153312/https://upload.wikimedia.org/wikipedia/commons/6/6d/A_teachers%27_manual_for_the_use_of_t

he_Binet-Simon_scale_of_intelligence_%28IA_teachersmanualfo00schw%29.pdf

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The glasgow norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, *51*(3), 1258–1270.

Sedoc, J., & Ungar, L. (2019). The Role of Protected Class Word Lists in Bias Identification of Contextualized Word Representations. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 55–61. https://doi.org/10.18653/v1/W19-3808

Seshadri, P., Pezeshkpour, P., & Singh, S. (2022). Quantifying social biases using templates is unreliable. *arXiv preprint arXiv:2210.04337*.

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493–498.

Shinar, E. H. (1975). Sexual stereotypes of occupations. *Journal of Vocational Behavior*, *7*(1), 99–111.

Shorrocks, R. (2018). Cohort change in political gender gaps in europe and canada: The role of modernization. *Politics & Society*, *46*(2), 135–175.

Simonsohn, U. (2017). The funnel plot is invalid because of this crazy assumption: R(n,d)=0 [Accessed: 2025-11-29]. https://web.archive.org/web/20251129195434/https://datacolada.org/58

Simonsohn, U., Montealegre, A., & Evangelidis, I. (2025). Stimulus sampling reimagined: Designing experiments with mix-and-match, analyzing results with stimulus plots. *Journal of Personality and Social Psychology*.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174.

Snyderman, M., & Herrnstein, R. (1983). Intelligence tests and the immigration act of 1924. *American Psychologist*, *38*(9), 986.

Spruyt, A., De Houwer, J., & Hermans, D. (2009). Modulation of automatic semantic priming by feature-specific attention allocation. *Journal of Memory and Language*, *61*(1), 37–54.

Spruyt, A., Gast, A., & Moors, A. (2011). The sequential priming paradigm: A primer. *Cognitive methods in social psychology*, 48–77.

Stanczak, K., & Augenstein, I. (2021). A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating Gender Bias in Machine Translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1679–1684. https://doi.org/10.18653/v1/P19-1164

Statistics Netherlands (Centraal Bureau voor de Statistiek). (2024). An increasing number of dutch people have completed higher education [News release; archived at: https://web.archive.org/web/20260625125508/https://www

.cbs.nl/en-gb/news/2024/41/an-increasing-number-of-dutch-people-have-completed-higher-education].

Stroud, N. J., Feldman, L., Wojcieszak, M., & Bimber, B. (2019). The consequences of forced versus selected political media exposure. *Human Communication Research*, *45*(1), 27–51.

Talat, Z., Névéol, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., Luccioni, S., Masoud, M., Mitchell, M., Radev, D., Sharma, S., Subramonian, A., Tae, J., Tan, S., Tunuguntla, D., & van der Wal, O. (2022). You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, 26–41. https://doi.org/10.18653/v1/2022.bigscience-1.3

Thau, M. (2019). How political parties use group-based appeals: Evidence from britain 1964–2015. *Political Studies*, *67*(1), 63–82.

Tontodimamma, A., Fontanella, L., Anzani, S., & Basile, V. (2022). An Italian lexical resource for incivility detection in online discourses. *Quality & Quantity*, *56*. https://doi.org/10.1007/s11135-022-01494-7

Trapnell, P. D., & Paulhus, D. L. (2012). Agentic and communal values: Their scope and measurement. *Journal of Personality Assessment*, *94*(1), 39–52.

Tsamadi, D., Falbén, J. K., Persson, L. M., Golubickis, M., Caughey, S., Sahin, B., & Macrae, C. N. (2020). Stereotype-based priming without stereotype activation: A tale of two priming tasks. *Quarterly Journal of Experimental Psychology*, *73*(11), 1939–1948.

Turpin, M., Michael, J., Perez, E., & Bowman, S. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, *36*, 74952–74965.

Uittenhove, K., Jeanneret, S., & Vergauwe, E. (2023). From lab-testing to web-testing in cognitive research: Who you test is more important than how you test. *Journal of cognition*, *6*(1), 13–13.

United States Congress. (1907). An act to regulate the immigration of aliens into the united states. https://govtrackus.s3.amazonaws.com/legislink/pdf/stat/34/STATUTE-34-Pg898.pdf

U.S. Bureau of Labor Statistics. (2024). Employed persons by detailed occupation and sex, 2023 annual averages [Current Population Survey (CPS), Table 11].

van Maanen, L., Katsimpokis, D., & van Campen, A. D. (2019). Fast and slow errors: Logistic regression to identify patterns in accuracy–response time relationships. *Behavior Research Methods*, *51*(5), 2378–2389.

van der Wal, O., Jumelet, J., Schulz, K., & Zuidema, W. (2022). The Birth of Bias: A case study on the evolution of gender bias in an English language model. *Proceedings of the 4th Workshop on Gender Bias in Natural Language*

Processing (GeBNLP), 75–75. https://doi.org/10.18653/v1/2022.gebnlp-1.8

van der Wal, O., Bachmann, D., Leidinger, A., van Maanen, L., Zuidema, W., & Schulz, K. (2024). Undesirable biases in NLP: Addressing challenges of measurement. *Journal of Artificial Intelligence Research*, *79*, 1–40.

Vania, C., Htut, P. M., Huang, W., Mungra, D., Yuanzhe Pang, R., Phang, J., Liu, H., Cho, K., & Bowman, S. R. (2021). Comparing test sets with Item Response Theory. https://arxiv.org/abs/2106.00840

Većkalov, B., van Stekelenburg, A., van Harreveld, F., & Rutjens, B. T. (2023). Who is skeptical about scientific innovation? examining worldview predictors of artificial intelligence, nanotechnology, and human gene editing attitudes. *Science Communication*, *45*(3), 337–366.

Verma, S., Vieweg, S., Corvey, W., Palen, L., Martin, J., Palmer, M., Schram, A., & Anderson, K. (2011). Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency [Number: 1]. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 385–392. https://doi.org/10.1609/icwsm.v5i1.14119

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (2020). Investigating Gender Bias in Language Models Using Causal Mediation Analysis. *Advances in Neural Information Processing Systems*, *33*, 12388–12401. https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html

Vigoda-Gadot, E., & Yuval, F. (2003). Managerial quality, administrative performance and trust in governance revisited: A follow-up study of causality. *International Journal of Public Sector Management*, *16*(7), 502–522.

von der Heyde, L., Haensch, A.-C., & Wenz, A. (2026). Vox populi, Vox AI? Using Large Language Models to estimate German vote choice. *Social Science Computer Review*, *44*(3), 549–571.

Voss, A., Rothermund, K., Gast, A., & Wentura, D. (2013). Cognitive processes in associative and categorical priming: A diffusion model analysis. *Journal of Experimental Psychology: General*, *142*(2), 536.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., et al. (2018). Bayesian inference for psychology. part ii: Example applications with jasp. *Psychonomic bulletin & review*, *25*(1), 58–76.

Wagner, C., Graells-Garrido, E., Garcia, D., & Menczer, F. (2016). Women through the glass ceiling: Gender asymmetries in Wikipedia. *EPJ Data Science*, *5*(1), 5. https://doi.org/10.1140/epjds/s13688-016-0066-4

Walter, T., Kirschner, C., Eger, S., Glavaš, G., Lauscher, A., & Ponzetto, S. P. (2021). Diachronic Analysis of German Parliamentary Proceedings: Ideological Shifts through the Lens of Political Biases. *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 51–60. https://doi.org/10.1109/JCDL52503.2021.00017

Wang, P., Yang, Y.-P., Tan, C.-H., Chen, Q.-W., & Cantfort, T. (2017). Gender stereotype activation versus lexical semantic activation: An erp study. *The Journal of General Psychology*, *144*(4), 283–308.

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., et al. (2018). Clinical information extraction applications: A literature review. *Journal of biomedical informatics*, *77*, 34–49.

Wang, Z., Wu, Z., Guan, X., Thaler, M., Koshiyama, A., Lu, S., Beepath, S., Ertekin, E., & Perez-Ortiz, M. (2024). Jobfair: A framework for benchmarking gender hiring bias in large language models. *Findings of the association for computational linguistics: EMNLP 2024*, 3227–3246.

Warm, T. A. (1978). *A Primer of Item Response Theory* (tech. rep.). Coast Guard Washington DC. https://files.eric.ed.gov/fulltext/ED171730.pdf

Warrens, M. J. (2015). On cronbach's alpha as the mean of all split-half reliabilities. *Quantitative Psychology Research: The 78th Annual Meeting of the Psychometric Society*, 293–300.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, *45*(4), 1191–1207.

Way, A. (2018). Quality expectations of machine translation. In *Translation quality assessment* (pp. 159–178). Springer.

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Chi, E., & Petrov, S. (2021). Measuring and Reducing Gendered Correlations in Pre-trained Models. https://doi.org/10.48550/arXiv.2010.06032

Weinberg, L. (2022). Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches. *Journal of Artificial Intelligence Research*, *74*, 75–109. https://doi.org/10.1613/jair.1.13196

Weisberg, M. (2007). Three kinds of idealization. *The Journal of Philosophy*, *104*(12), 639–659.

Wentura, D., & Degner, J. (2010). A practical guide to sequential priming and related tasks. *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*, 95–116.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. https://arxiv.org/abs/2302.11382

White, K. R., Danek, R. H., Herring, D. R., Taylor, J. H., & Crites, S. L. (2018). Taking priming to task. *Social Psychology*.

White, M. J., Kruczek, T. A., Brown, M. T., & White, G. B. (1989). Occupational sex stereotypes among college students. *Journal of Vocational Behavior*, *34*(3), 289–298.

Whitlock, M., & Schluter, D. (2015). *The analysis of biological data* (Vol. 768). Roberts Publishers.

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *elife*, *8*, e49547.

Winham, S. J., de Andrade, M., & Miller, V. M. (2015). Genetics of cardiovascular disease: Importance of sex and ethnicity. *Atherosclerosis*, *241*(1), 219–228. https://doi.org/10.1016/j.atherosclerosis.2015.03.021

Wong, K., & Paritosh, P. (2022). K-rater reliability: The correct unit of reliability for aggregated human annotations. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 378–384.

Wong, K., Paritosh, P., & Aroyo, L. (2021). Cross-replication Reliability - An Empirical Approach to Interpreting Inter-rater Reliability. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 7053–7065. https://doi.org/10.18653/v1/2021.acl-long.548

Wu, M., Tam, H. P., & Jen, T.-H. (2016). Differential item function. In *Educational Measurement for Applied Researchers: Theory into Practice* (pp. 207–225). Springer Singapore. https://doi.org/10.1007/978-981-10-3302-5_11

Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., Abdulnour, R.-E. E., et al. (2024). Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: A model evaluation study. *The Lancet Digital Health*, *6*(1), e12–e22.

Zaghouani, W. (2026). Position: What are we measuring? rethinking evaluation in natural language generation. *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, 1021–1028.

Zeinert, P., Inie, N., & Derczynski, L. (2021). Annotating Online Misogyny. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3181–3197. https://doi.org/10.18653/v1/2021.acl-long.247

Zhang, H., Sneyd, A., & Stevenson, M. (2020). Robustness and Reliability of Gender Bias Assessment in Word Embeddings: The Role of Base Pairs. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 759–769. https://aclanthology.org/2020.aacl-main.76

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. (2022). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, X., Li, Q., Sun, S., & Zuo, B. (2018). The time course from gender categorization to gender-stereotype activation. *Social Neuroscience*, *13*(1), 52–60.

Zhang, Y., Zhang, Y., Halpern, B. M., Patel, T., & Scharenborg, O. (2022). Mitigating bias against non-native accents. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2022*, 3168–3172.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K.-W. (2019). Gender Bias in Contextualized Word Embeddings. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 629–634. https://doi.org/10.18653/v1/N19-1064

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20. https://doi.org/10.18653/v1/N18-2003

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). How does nlp benefit legal system: A summary of legal artificial intelligence. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5218–5230.

Zmerli, S., & Newton, K. (2008). Social trust and attitudes toward democracy. *Public opinion quarterly*, *72*(4), 706–724.

Samenvatting

Towards Sound Measurement of Stereotypes and Attitudes in Humans and Language Models – van Dominik Bachmann (Naar een verantwoord meten van stereotypen en attitudes bij mensen en taalmodellen)

Een groot deel van de empirische wetenschap draait om meten: het comprimeren van de bijna oneindige complexiteit van de buitenwereld tot een eindige set gegevens (meestal in de vorm van getallen). Meten kan grote gevolgen hebben: op basis van gegevens worden beslissingen (bijvoorbeeld diagnoses) genomen en conclusies getrokken over de wereld. Deze beslissingen en conclusies zijn slechts valide voor zover de gegevens betekenen wat men denkt dat ze betekenen. Gezien het belang van goede metingen in empirisch onderzoek, is het doel van dit proefschrift om bij te dragen aan de ontwikkeling van hoogwaardige meetinstrumenten voor mensen en taalmodellen.

Deel 1 van dit proefschrift slaat een brug tussen *Natural Language Processing* (NLP; oftewel het subgebied van kunstmatige intelligentie dat zich bezig houdt met natuurlijke taalverwerking) en Psychometrie (oftewel het subgebied van de psychologie dat zich bezighoudt met de theorie en praktijk van meten). We betogen hier dat psychometrie NLP voorziet van theoretische perspectieven (namelijk betrouwbaarheid en validiteit, hoofdstuk 2, en itemresponstheorie, hoofdstuk 3). Dit kan helpen bij het aanpakken van tekortkomingen van hedendaagse meetinstrumenten voor bias in NLP modellen (sommige modellen bijvoorbeeld gaan er onterecht vanuit dat sommige beroepen overwegend mannelijk zijn). Daarnaast bieden we, ter ondersteuning van de validatie van bias-meetinstrumenten voor gender, een grote dataset met menselijke beoordelingen van genderstereotypen in beroepen (hoofdstuk 4).

Deel 2 onderzoekt de theoretische aannames die ten grondslag liggen aan het meten van menselijke stereotypen en het voorspellen van meningen. In hoofdstuk 5 testen we specifiek of, zoals vaak wordt aangenomen, bewijs voor genderstereotypen kan worden verkregen met een sequentieel primingparadigma, wanneer deel-

nemers zich niet bewust zijn van wat we proberen te meten. Vervolgens bespreken we in hoofdstuk 6 een vooronderzoek met vragenlijsten, waarin we het krantenconsumptiegedrag en een aantal van de (politieke) meningen van mensen hebben gemeten. Dit vooronderzoek maakte deel uit van een project waarmee we wilden vaststellen in hoeverre grote taalmodellen kunnen voorspellen hoe verschillende groepen mensen reageren op verschillende vragenlijstitems (een benadering die 'silicon sampling' genoemd wordt).

Het hoofdargument van dit proefschrift is dat de aannames die ten grondslag liggen aan meetinstrumenten en het meetproces expliciet gemaakt, gecommuniceerd, onderzocht en getoetst zouden moeten worden.

Abstract

Towards Sound Measurement of Stereotypes and Attitudes in Humans and Language Models – by Dominik Bachmann

Much of empirical science involves measurement: compressing the near-infinite complexity of the outside world into a finite set of data (usually in the form of numbers). Measurement is potentially consequential: Based on data, decisions (e.g., diagnoses) are made, and conclusions are drawn about the world. These decisions and conclusions are valid only to the extent to which the data mean what one thinks they mean. Given the importance of good measurement in empirical research, the goal of this dissertation is to contribute to the development of high quality measurement tools for humans and language models.

Part 1 of this dissertation bridges Natural Language Processing (NLP; i.e., the subfield of artificial intelligence specialized in understanding and acting on human language) and Psychometrics (i.e., the subfield of psychology concerned with the theory and practice of measurement). Here, we argue that psychometrics provides NLP with theoretical lenses (i.e., reliability and validity, Chapter 2, and item response theory, Chapter 3) that can help address shortcomings of contemporary NLP bias measurement tools. Additionally, to support the validation of those NLP gender bias measurement tools that rely on occupational terms, we provide a large dataset of human ratings for occupational gender stereotypes (Chapter 4).

Part 2 explores theoretical assumptions that underlie an approach to measuring human stereotypes, and an approach to approximating human opinions. Specifically, in Chapter 5, we test whether, as is often assumed, evidence for gender stereotypes can be obtained with a sequential priming paradigm, when participants are unaware of what we attempt to measure. Subsequently, in Chapter 6, we discuss a pilot questionnaire study, in which we measured people's newspaper consumption and several of their (political) opinions. This pilot study was part of a project with which we aimed to assess the extent to which Large

Language Models can predict how different groups of people respond to different questionnaire items (i.e., an approach called silicon sampling).

The main argument of this dissertation is that assumptions that underlie measurement tools and one's process of measuring deserve to be made explicit, communicated, scrutinized, and tested.

Curriculum Vitae

Dominik Bachmann



Institute for Logic, Language & Computation,
University of Amsterdam

Supervisors Dr Katrin Schulz; Dr Leendert van Maanen
Email dominik.bachmann.psychology@gmail.com

Academic Background

PhD, Institute for Logic, Language & Computation, University of Amsterdam 2022–2026

Supervisors: Dr Katrin Schulz; Dr Leendert van Maanen (Utrecht University)

Guest researcher, Department of Experimental Psychology, Utrecht University 2022–2026

Dissertation Defence Nov 25, 2026

Research Master of Science in Psychology, University of Amsterdam 2019–2021

Major: Cognitive Psychology Minor: Social Psychology
GPA: 8.7 *(cum Laude)*

Thesis: *"What came first, the mind wandering or the task? Investigating the validity of the probe-caught method"* 2021
Supervisor: Dr Filip van Opstal, grade: 9

QHELP 2021 International Statistics Seminar on Measurement Theory 2021

International Bachelor of Science in Psychology, Leiden University 2015–2018

GPA: 9.0 *(Summa cum Laude)*, Valedictorian

Thesis: *"No evidence of enhanced schema updating after consolidation"* 2018
Supervisor: Dr Franka Richter, grade: 9

Exchange semester, UBC Vancouver local GPA: 94 2017

Honours College "Science and Society", Faculty of Social Sciences, Leiden University 2015–2018

Teaching Experience & Internships

Teaching as part of the PhD position, University of Amsterdam 2022–2025

"Cognition, Language and Communication" Bachelor: Designing Research, Statistics in R
Supervisor of Student Intervisions, Psychology Research Master
Psychology Bachelor: Addiction & Compulsive Disorders; Personality & Personality Disorders

***Docent* / Workgroup Teacher, UvA Psychology Bachelor:** 2022

Mood, Anxiety and Psychotic Disorders

Teaching Assistant, UvA Psychology Research Master: 2020–2021

Consciousness; Scientific Writing and Presenting

Volunteer, Consciousness Psychology Unit 2020–2021

Supervisor: Dr Timo Stein, University of Amsterdam

Research Master Internship: 2020

"Unconscious sound-symbolism effects reexamined", grade: 9

Supervisor: Dr Timo Stein, University of Amsterdam

Volunteer, Memory Psychology Unit 2019

Supervisor: Dr Franka Richter, Leiden University

Research Internship, Music Psychology Unit 2018–2019

Supervisor: Dr Rebecca Schaefer, Leiden University

Research Internship, Clinical Psychology Unit 2016

Supervisor: Dr Meriem Manaï, Leiden University

Titles in the ILLC Dissertation Series:

ILLC DS-2021-11: **Sirin Botan**
Strategyproof Social Choice for Restricted Domains

ILLC DS-2021-12: **Michael Cohen**
Dynamic Introspection

ILLC DS-2021-13: **Dazhu Li**
Formal Threads in the Social Fabric: Studies in the Logical Dynamics of Multi-Agent Interaction

ILLC DS-2021-14: **Álvaro Piedrafita**
On Span Programs and Quantum Algorithms

ILLC DS-2022-01: **Anna Bellomo**
Sums, Numbers and Infinity: Collections in Bolzano's Mathematics and Philosophy

ILLC DS-2022-02: **Jan Czajkowski**
Post-Quantum Security of Hash Functions

ILLC DS-2022-03: **Sonia Ramotowska**
Quantifying quantifier representations: Experimental studies, computational modeling, and individual differences

ILLC DS-2022-04: **Ruben Brokkelkamp**
How Close Does It Get?: From Near-Optimal Network Algorithms to Suboptimal Equilibrium Outcomes

ILLC DS-2022-05: **Lwenn Bussière-Carae**
No means No! Speech Acts in Conflict

ILLC DS-2022-06: **Emma Mojet**
Observing Disciplines: Data Practices In and Between Disciplines in the 19th and Early 20th Centuries

ILLC DS-2022-07: **Freek Gerrit Witteveen**
Quantum information theory and many-body physics

ILLC DS-2023-01: **Subhasree Patro**
Quantum Fine-Grained Complexity

ILLC DS-2023-02: **Arjan Cornelissen**
Quantum multivariate estimation and span program algorithms

ILLC DS-2023-03: **Robert Paßmann**
Logical Structure of Constructive Set Theories

ILLC DS-2023-04: **Samira Abnar**
Inductive Biases for Learning Natural Language

ILLC DS-2023-05: **Dean McHugh**
Causation and Modality: Models and Meanings

ILLC DS-2023-06: **Jialiang Yan**
Monotonicity in Intensional Contexts: Weakening and: Pragmatic Effects under Modals and Attitudes

ILLC DS-2023-07: **Yiyan Wang**
Collective Agency: From Philosophical and Logical Perspectives

ILLC DS-2023-08: **Lei Li**
Games, Boards and Play: A Logical Perspective

ILLC DS-2023-09: **Simon Rey**
Variations on Participatory Budgeting

ILLC DS-2023-10: **Mario Giulianelli**
Neural Models of Language Use: Studies of Language Comprehension and Production in Context

ILLC DS-2023-11: **Guillermo Menéndez Turata**
Cyclic Proof Systems for Modal Fixpoint Logics

ILLC DS-2023-12: **Ned J.H. Wontner**
Views From a Peak: Generalisations and Descriptive Set Theory

ILLC DS-2024-01: **Jan Rooduijn**
Fragments and Frame Classes: Towards a Uniform Proof Theory for Modal Fixed Point Logics

ILLC DS-2024-02: **Bas Cornelissen**
Measuring musics: Notes on modes, motifs, and melodies

ILLC DS-2024-03: **Nicola De Cao**
Entity Centric Neural Models for Natural Language Processing

ILLC DS-2024-04: **Ece Takmaz**
Visual and Linguistic Processes in Deep Neural Networks: A Cognitive Perspective

ILLC DS-2024-05: **Fatemeh Seifan**
Coalgebraic fixpoint logic Expressivity and completeness result

ILLC DS-2024-06: **Jana Sotáková**
Isogenies and Cryptography

ILLC DS-2024-07: **Marco Degano**
Indefinites and their values

ILLC DS-2024-08: **Philip Verduyn Lunel**
Quantum Position Verification: Loss-tolerant Protocols and Fundamental Limits

ILLC DS-2024-09: **Rene Allerstorfer**
Position-based Quantum Cryptography: From Theory towards Practice

ILLC DS-2024-10: **Willem Feijen**
Fast, Right, or Best? Algorithms for Practical Optimization Problems

ILLC DS-2024-11: **Daira Pinto Prieto**
Combining Uncertain Evidence: Logic and Complexity

ILLC DS-2024-12: **Yanlin Chen**
On Quantum Algorithms and Limitations for Convex Optimization and Lattice Problems

ILLC DS-2024-13: **Jaap Jumelet**
Finding Structure in Language Models

ILLC DS-2025-01: **Julian Chingoma**
On Proportionality in Complex Domains

ILLC DS-2025-02: **Dmitry Grinko**
Mixed Schur-Weyl duality in quantum information

ILLC DS-2025-03: **Rochelle Choenni**
Multilinguality and Multiculturalism: Towards more Effective and Inclusive Neural Language Models

ILLC DS-2025-04: **Aleksi Anttila**
Not Nothing: Nonemptiness in Team Semantics

ILLC DS-2025-05: **Niels M. P. Neumann**
Adaptive Quantum Computers: decoding and state preparation

ILLC DS-2025-06: **Alina Leidinger**
Towards Language Models that benefit us all: Studies on stereotypes, robustness, and values

ILLC DS-2025-07: **Zhi Zhang**
Advancing Vision and Language Models through Commonsense Knowledge, Efficient Adaptation and Transparency

ILLC DS-2025-08: **Sophie Klumper**
The Gap and the Gain: Improving the Approximate Mechanism Design Frontier in Constrained Environments

ILLC DS-2025-09: **Jordi Rudo Weggemans**
Quantum versus classical resources in computational complexity

ILLC DS-2026-01: **Bryan Eikema**
A Sampling-Based Exploration of Neural Text Generation Models

ILLC DS-2026-02: **Marten Folkertsma**
Empowering Quantum Computation with: Measurements, Catalysts, and Guiding States

ILLC DS-2026-03: **Valentin Richard**
Presuppositional and Dynamic Aspects of Questions

ILLC DS-2026-04: **Puyu Yang**
Bringing Science to the Public: The Role of Wikipedia in Scientific Communication

ILLC DS-2026-05: **Johannes Kloibhofer**
Cycles with Annotations: Non-Wellfounded Proof Theory of Modal Fixpoint Logics

ILLC DS-2026-06: **Danish Kashaev**
Approximation via duality in offline, online and strategic settings

ILLC DS-2026-07: **Oskar van der Wal**
Taking a Step Back: Measuring and Mitigating Bias in Language Models

ILLC DS-2026-08: **Lynn Engelberts**
Classical and Quantum Cryptanalysis of Lattices and Codes

ILLC DS-2026-09: **Xiaoyu Tong**
Dissecting Incongruity: Metaphor and Humor Understanding of Large Language Models

ILLC DS-2026-10: **Qian Chen**
Lattices of Tense Logics: Tabularity, Completeness and Decidability

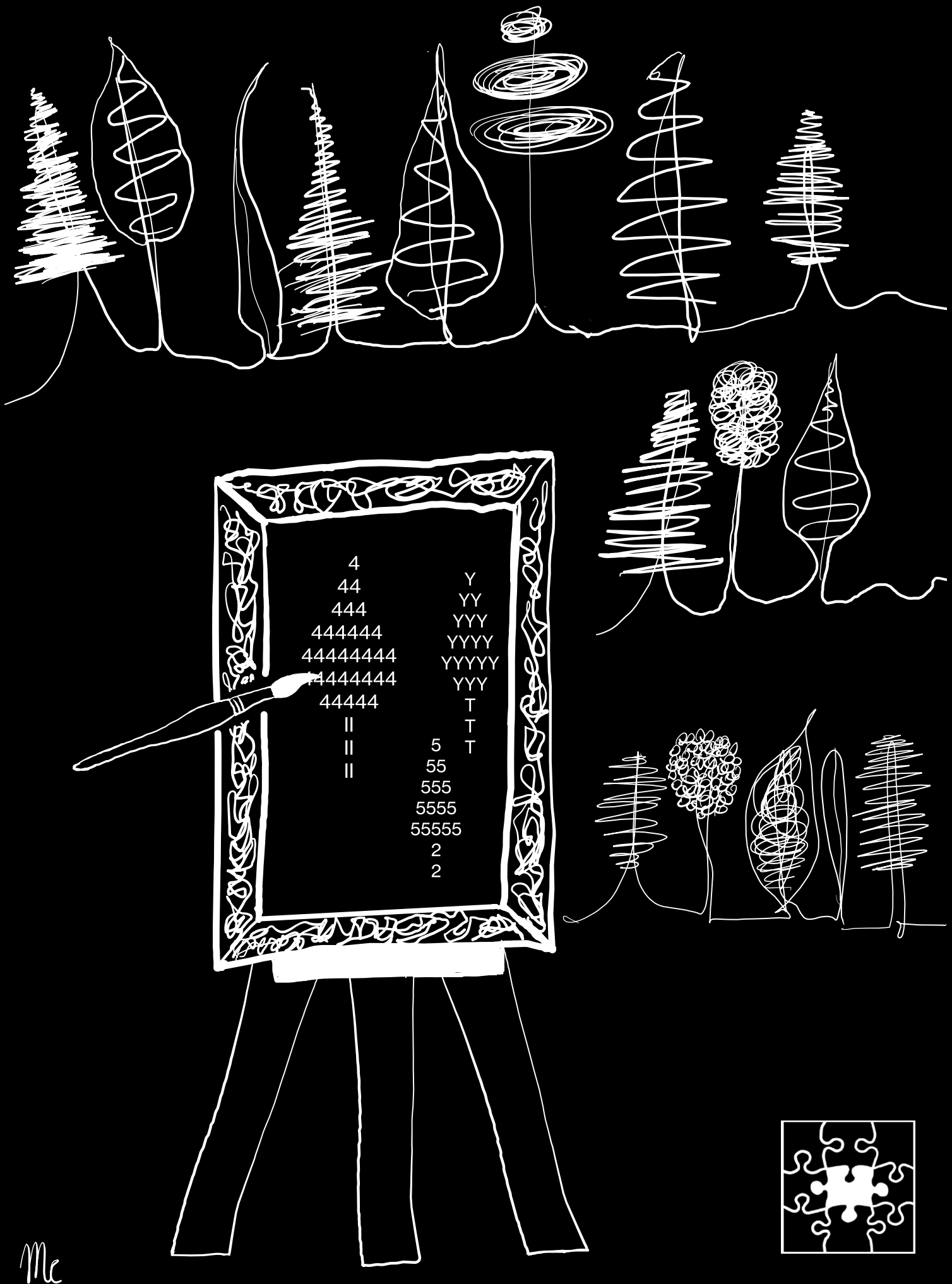
ILLC DS-2026-11: **Daniil Khaitovich**
From arbitrary sets to social groups: group agency in multi-agent modal logic

ILLC DS-2026-12: **Eline de Jong**
Technological Understanding and Ethics Readiness of Technology: the role of understanding and timing in the ethics of quantum technologies

ILLC DS-2026-13: **Sophie Nagler**
Measuring Meaning via Inference-Behaviour Semantics: A Substructural Approach to Proof-Theoretic Semantics

ILLC DS-2026-14: **Rodrigo Nicolau Almeida**
Representations and model theory of free Heyting algebras

ILLC DS-2026-15: **Dominik Bachmann**
Towards Sound Measurement of Stereotypes and Attitudes in Humans and Language Models


4
44
444
444444
44444444
44444444
44444
II
II
II
Y
YY
YYY
YYYY
YYYYY
YYY
T
T
T
5
55
555
5555
55555
2
2