Undesirable Biases in NLP

Addressing Challenges of Measurement

van der Wal, O.; Bachmann, D.; Leidinger, A.; van Maanen, L.; Zuidema, W.; Schulz, K.

DOI
10.1613/jair.1.15195

Publication date
2024

Document Version
Final published version

Published in
Journal of Artificial Intelligence Research

Journal of Artificial Intelligence Research 79 (2024) 1-40 Submitted 07/2023; published 01/2024

Undesirable Biases in NLP: Addressing Challenges of Measurement

Oskar van der Wal O.D.VANDERWAL@UVA.NL
Institute for Logic, Language and Computation, University of Amsterdam

Dominik Bachmann D.BACHMANN@UVA.NL
Institute for Logic, Language and Computation, University of Amsterdam
Department of Experimental Psychology, Utrecht University

Alina Leidinger A.J.LEIDINGER@UVA.NL
Institute for Logic, Language and Computation, University of Amsterdam

Leendert van Maanen L.VANMAANEN@UU.NL
Department of Experimental Psychology, Utrecht University

Willem Zuidema W.H.ZUIDEMA@UVA.NL
Katrin Schulz K.SCHULZ@UVA.NL
Institute for Logic, Language and Computation, University of Amsterdam

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.

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), jobseekers 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., 2021).

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 et al., 2022b), 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 endevor 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). We then discuss reliability (Section 3) and construct validity (Section 4), and the use of these concepts when evaluating bias measures in an NLP context. Section 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 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. 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 2.1 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). 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, 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 2.4) 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 2.3, we think that there is a lot of promise in connecting NLP research to work in psychometrics. The rest of Section 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 2.4 and the notions of validity and reliability in Section 2.5.

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 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 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., 2020; 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 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 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

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., 2020). 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 1: Examples of NLP benchmarks that operationalize (gender) bias through contrasting sets of words or sentences.

example, Natural Language Inference). Table 2 provides some examples of prompts that are used to make a language model respond to items from benchmark datasets.

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-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

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: 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 1). Prompts stem from Scao et al. (2022) and PromptSource (Bach et al., 2022), respectively.

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 the issue of bias in NLP models.

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

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!".

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.

2.4 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

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).

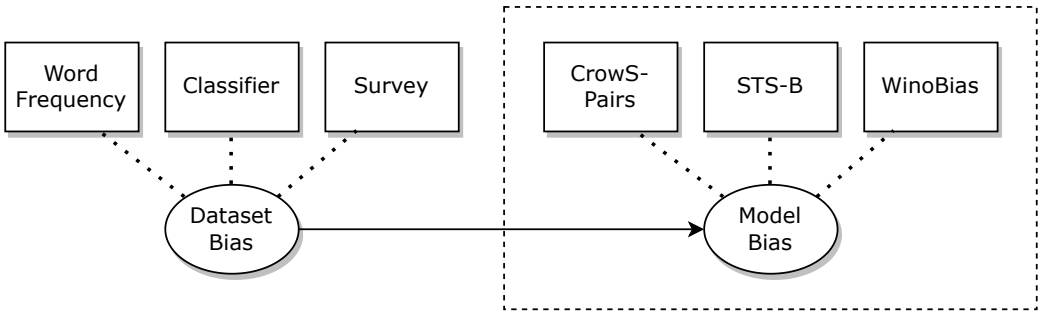


Figure 1: We assume that a training dataset's bias influences the bias of a model trained on that data (but other possible sources of bias are possible, e.g., model compression may amplify existing biases (Hooker et al., 2020)). Training dataset bias and model bias are unobservable constructs (circle) that both have different possible operationalizations (squares).

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 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 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., 2020), and WinoBias (Zhao et al., 2018) are examples of model bias operationalizations.

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).

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]

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 3) and construct validity (Section 4) of bias measures in more detail and provide strategies for assessing these in an NLP setting.

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

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 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

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 3: Examples of the reliability types we discuss in Section 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.

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) 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 3 provides an overview of these discussed subtypes and example applications.

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

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.

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 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 1). Authors like Wong and colleagues (e.g., Wong et al., 2021; Wong & Paritosh, 2022) 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.

3.2 Internal Consistency

Internal consistency can be relevant for evaluating the quality of bias measures based on different items (see Section 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 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

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).

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 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 4.3.

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 2.1) 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 et al., 2022a), StereoSet (Zhang et al., 2022a), 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). 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.

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).

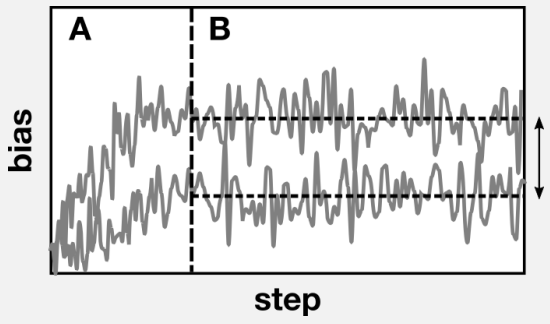
Time-Based Test-Retest Reliability Finally, a way of retaining test-retest reliability's temporal component could be to compare bias measurements for models trained on data from the same corpus but collected at different points in time — for instance, datasets extracted from the same social media platform 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).



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 4.1), divergent validity (4.2), and content validity (4.3). Table 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.

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). 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 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

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 4: An overview of the types of construct validity we discuss in Section 4. Examples are given in the last column.

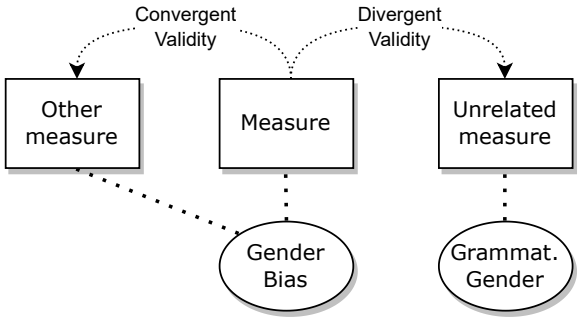


Figure 2: This figure illustrates the difference between convergent and divergent validity (see Section 4.2). In this example, the convergent validity is assessed by testing how related a gender bias measure is to another gender bias measure. The divergent validity, instead, is assessed by testing whether the gender bias measure is not strongly correlated with a measure for another, but easily confounded construct (e.g., grammatical gender).

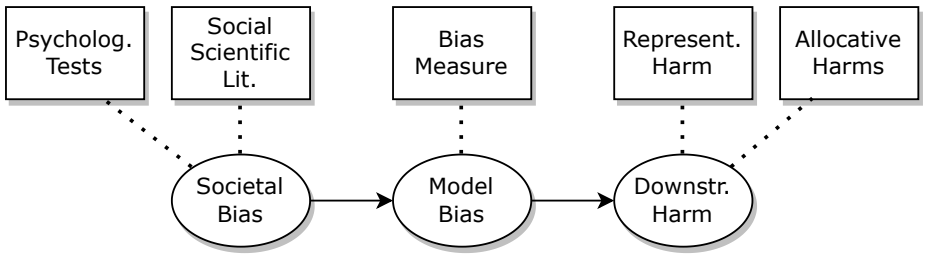


Figure 3: In Section 4.1, we discuss two ways to validate bias measures through related concepts: (1) testing whether the found bias reflects pre-existing stereotypes in society (e.g., informed by psychological tests or the social scientific literature), and (2) testing the relationship of the bias to 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.

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 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 3).

As NLP technologies model regularities in natural language, bias measures should for example assign higher bias values to words associated with (human) 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., 2020; 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., 2023; Nosek et al., 2015). If the external criterion based on which a bias measure is validated is itself not valid, the validity of the

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).

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 *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.

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). 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): 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.

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]

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.
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 helpfull 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.

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 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.

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.

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

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.
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 2.3) 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.

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 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.

5. From Theory to Practice: Designing Good Bias Measures

How do we put the lesson 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.

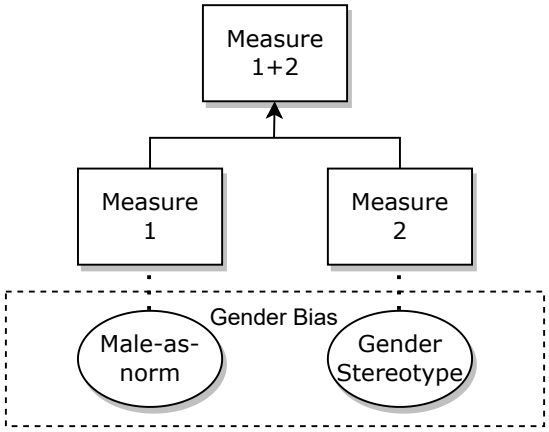


Figure 4: **Content validity:** in this example, *male-as-norm bias* and *gender stereotypes* are hypothesized to be separate subconcepts of the model's construct gender bias. The *male-as-norm* bias reflects the idea that the male gender is assumed as default, unless explicitly indicated otherwise (Danesi, 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).

(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)?
- For which (downstream) tasks do we develop this bias measure (§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 (§3) and valid (§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 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 (§3.4) of our measure)?
- What are our assumptions about the bias construct (§2.4)? What (potential) subcomponents of model bias are relevant for this bias measure (§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 (§4.1) and divergent validity (§4.2).
- If our measure depends on downstream task performance (e.g., WinoBias, see Table 1): How do we expect the model's bias to influence its behavior on the downstream task (§4.2)?
- What assumptions do we make during the "translation step" (§2.3) of psychometric concepts to the NLP context? What are the theoretical consequences of making these assumptions? What definitions of reliability (§3) and validity (§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 (§3) and validity (§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 (§3.3)?
- Do human annotators judge candidate items for our bias measure? If so, what is the inter-rater reliability of their ratings (§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 (§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* (§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) (§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 (§4.1)?
- Are there relevant measures for testing the *divergent validity* (§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 (§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)?
- Could our measure accommodate subcategorizations of bias (§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?

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).

(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?

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 and 4.1). Another survey by Blodgett et al. (2021) — but of measurement tools based on constrastive 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 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 scientists. 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.

7. Conclusions

Bias in NLP is an 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.

Acknowledgments

The two first authors, Oskar van der Wal and Dominik Bachmann, contributed equally to the paper. We thank our four anonymous reviewers for their thoughtful and elaborate feedback.

This publication is part of the project "The biased reality of online media - Using stereotypes to make media manipulation visible" (with project number 406.DI.19.059) of the research programme Open Competition Digitalisation-SSH, which is financed by the Dutch Research Council (NWO).

References

Abbasi, A., Dobolyi, D., Lalor, J. P., Netemeyer, R. G., Smith, K., & Yang, Y. (2021). Constructing a Psychometric Testbed for Fair Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3748–3758, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 551–564.

Akyürek, A. F., Kocyigit, M. Y., Paik, S., & Wijaya, D. T. (2022). Challenges in Measuring Bias via Open-Ended Language Generation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 76–76, Seattle, Washington. Association for Computational Linguistics.

Amidei, J., Piwek, P., & Willis, A. (2020). Identifying Annotator Bias: A new IRT-based method for bias identification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4787–4797, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Antoniak, M., & Mimno, D. (2021). Bad seeds: Evaluating lexical methods for bias measurement. In *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)*, pp. 1889–1904.

Aribandi, V., Tay, Y., & Metzler, D. (2021). How Reliable are Model Diagnostics?. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1778–1785, Online. Association for Computational Linguistics.

Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, *34*(4), 555–596.

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. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 93–104.

Balayn, A., & Gürses, S. (2021). Beyond debiasing: Regulating ai and its inequalities. *EDRi Report*.

Barocas, S., Crawford, K., Shapiro, A., & Wallach, H. (2017). The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual conference of the special interest group for computing, information and society*.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA. Association for Computing Machinery. tex.ids= bender2021DangersStochasticParrotsa.

Bhowmick, P. K., Basu, A., & Mitra, P. (2008). An Agreement Measure for Determining Inter-Annotator Reliability of Human Judgements on Affective Text. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pp. 58–65, Manchester, UK. Coling 2008 Organizing Committee.

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. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Vol. 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online. Association for Computational Linguistics.

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., & Wallach, H. (2021). Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *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)*, pp. 1004–1015, Online. Association for Computational Linguistics.

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. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bommasani, R., & Liang, P. (2022). Trustworthy social bias measurement. *arXiv preprint arXiv:2212.11672*.

Bordia, S., & Bowman, S. R. (2019). Identifying and Reducing Gender Bias in Word-Level Language Models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061–1071. Place: US Publisher: American Psychological Association.

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., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901. Curran Associates, Inc.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

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. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 561–570, Dublin, Ireland. Association for Computational Linguistics.

Cheng, L., Varshney, K. R., & Liu, H. (2021). Socially Responsible AI Algorithms: Issues, Purposes, and Challenges. *Journal of Artificial Intelligence Research*, *71*, 1137–1181.

Ciora, C., Iren, N., & Alikhani, M. (2021). Examining covert gender bias: A case study in turkish and english machine translation models. In *Proceedings of the 14th International Conference on Natural Language Generation*, pp. 55–63.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. Place: US Publisher: American Psychological Association.

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., Mitani, A., Montanari, A., Nado, Z., Natarajan, V., Nielson, C., Osborne, T. F., Raman, R., Ramasamy, K., Sayres, R., Schrouff, J., Seneviratne, M., Sequeira, S., Suresh, H., Veitch, V., Vladymyrov, M., Wang, X., Webster, K., Yadlowsky, S., Yun, T., Zhai, X., & Sculley, D. (2022). Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research*, *23*(226), 1–61.

Danesi, M. (2014). *Dictionary of media and communications*. Routledge.

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. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 120–128, New York, NY, USA. Association for Computing Machinery.

De Cao, N., Schmid, L., Hupkes, D., & Titov, I. (2022). Sparse Interventions in Language Models with Differentiable Masking. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 16–27, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1693–1706, Seattle, United States. Association for Computational Linguistics.

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. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pp. 246–267, Online only. Association for Computational Linguistics.

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., & Williams, A. (2020). Multi-Dimensional Gender Bias Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 314–331, Online. Association for Computational Linguistics.

Du, Y., Fang, Q., & Nguyen, D. (2021). Assessing the Reliability of Word Embedding Gender Bias Measures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10012–10034, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. *American Psychologist*, *75*, 301–315. place: US publisher: American Psychological Association.

Ethayarajh, K. (2020). Is Your Classifier Actually Biased? Measuring Fairness under Uncertainty with Bernstein Bounds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2914–2919, Online. Association for Computational Linguistics.

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). Understanding Undesirable Word Embedding Associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1696–1705, Florence, Italy. Association for Computational Linguistics.

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.

Field, A., Blodgett, S. L., Waseem, Z., & Tsvetkov, Y. (2021). A Survey of Race, Racism, and Anti-Racism in NLP. In *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)*, pp. 1905–1925, Online. Association for Computational Linguistics.

Field, A., & Tsvetkov, Y. (2020). Unsupervised Discovery of Implicit Gender Bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 596–608, Online. Association for Computational Linguistics.

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. In *Proceedings of the international AAAI conference on web and social media*, Vol. 12.

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.

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.

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. In *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)*, pp. 1926–1940, Online. Association for Computational Linguistics.

Goldfarb-Tarrant, S., Ungless, E., Balkir, E., & Blodgett, S. L. (2023). This prompt is measuring <MASK>: Evaluating bias evaluation in language models. *arXiv preprint arXiv:2305.12757*.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *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)*, pp. 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Gonen, H., Ravfogel, S., & Goldberg, Y. (2022). Analyzing Gender Representation in Multilingual Models. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pp. 67–77, Dublin, Ireland. Association for Computational Linguistics.

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.

Journal of Personality and Social Psychology, *97*(1), 17–41. Place: US Publisher: American Psychological Association.

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.

Harrington, D. (2009). *Confirmatory Factor Analysis*. Oxford University Press, USA.

Hogenboom, S. A. M., Schulz, K., & Van Maanen, L. (2023). Implicit association tests: Stimuli validation from participant responses. Forthcoming.

Hooker, S., Moorosi, N., Clark, G., Bengio, S., & Denton, E. (2020). Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5491–5501, Online. Association for Computational Linguistics.

Jacobs, A. Z., & Wallach, H. (2021). Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 375–385, New York, NY, USA. Association for Computing Machinery.

Jiao, M., & Luo, Z. (2021). Gender Bias Hidden Behind Chinese Word Embeddings: The Case of Chinese Adjectives. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pp. 8–15, Online. Association for Computational Linguistics.

Kidd, C., & Birhane, A. (2023). How AI can distort human beliefs. *Science*, *380*(6651), 1222–1223.

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.

Kline, P. (2014). *An easy guide to factor analysis*. Routledge.

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. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, Vol. 2016, p. 648. NIH Public Access.

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.

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*.

Limisiewicz, T., & Mareček, D. (2022). Don't Forget About Pronouns: Removing Gender Bias in Language Models Without Losing Factual Gender Information. In *Proceedings*

of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pp. 17–29, Seattle, Washington. Association for Computational Linguistics.

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.

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. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Vol. 202 of *Proceedings of Machine Learning Research*, pp. 22631–22648. PMLR.

Malik, V., Dev, S., Nishi, A., Peng, N., & Chang, K.-W. (2022). Socially Aware Bias Measurements for Hindi Language Representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1041–1052, Seattle, United States. Association for Computational Linguistics.

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.

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). On Measuring Social Biases in Sentence Encoders. In *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)*, pp. 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Meade, N., Poole-Dayan, E., & Reddy, S. (2022). An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

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.

Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In *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)*, pp. 5356–5371, Online. Association for Computational Linguistics.

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967, Online. Association for Computational Linguistics.

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, *18*, 301–319. Place: US Publisher: American Psychological Association.

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., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E. L., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E. J., Wilson, R., & Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. Publisher: American Association for the Advancement of Science.

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. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Orgad, H., & Belinkov, Y. (2022). Choose your lenses: Flaws in gender bias evaluation. *GeBNLP 2022*, 151.

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.

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. In *ICLR 2022-Tenth International Conference on Learning Representations*.

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.

Sedoc, J., & Ungar, L. (2019). The Role of Protected Class Word Lists in Bias Identification of Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pp. 55–61, Florence, Italy. Association for Computational Linguistics.

Seshadri, P., Pezeshkpour, P., & Singh, S. (2022). Quantifying social biases using templates is unreliable. *arXiv preprint arXiv:2210.04337*.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1679–1684.

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. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 26–41, virtual+Dublin. Association for Computational Linguistics.

Tontodimamma, A., Fontanella, L., Anzani, S., & Basile, V. (2022). An Italian lexical resource for incivility detection in online discourses. *Quality & Quantity*, *56*.

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. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 385–392. Number: 1.

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. In *Advances in Neural Information Processing Systems*, Vol. 33, pp. 12388–12401. Curran Associates, Inc.

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.

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. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 51–60.

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. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pp. 75–75, Seattle, Washington. Association for Computational Linguistics.

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.

Warrens, M. J. (2015). On cronbach's alpha as the mean of all split-half reliabilities. In *Quantitative Psychology Research: The 78th Annual Meeting of the Psychometric Society*, pp. 293–300. Springer.

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. (2020). Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint 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.

Whitlock, M., & Schluter, D. (2015). *The analysis of biological data*, Vol. 768. Roberts Publishers.

Winham, S. J., de Andrade, M., & Miller, V. M. (2015). Genetics of cardiovascular disease: Importance of sex and ethnicity. *Atherosclerosis*, *241*(1), 219–228.

Wong, K., & Paritosh, P. (2022). k-rater reliability: The correct unit of reliability for aggregated human annotations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 378–384.

Wong, K., Paritosh, P., & Aroyo, L. (2021). Cross-replication Reliability - An Empirical Approach to Interpreting Inter-rater Reliability. In *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)*, pp. 7053–7065, Online. Association for Computational Linguistics.

Zeinert, P., Inie, N., & Derczynski, L. (2021). Annotating Online Misogyny. In *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)*, pp. 3181–3197, Online. Association for Computational Linguistics.

Zhang, H., Sneyd, A., & Stevenson, M. (2020). Robustness and Reliability of Gender Bias Assessment in Word Embeddings: The Role of Base Pairs. In *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*, pp. 759–769, Suzhou, China. Association for Computational Linguistics.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. (2022a). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, Y., Zhang, Y., Halpern, B. M., Patel, T., & Scharenborg, O. (2022b). Mitigating bias against non-native accents. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Vol. 2022, pp. 3168–3172.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K.-W. (2019). Gender Bias in Contextualized Word Embeddings. In *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)*, pp. 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5218–5230.