Measurement Modeling for Bias Metrics for LMs

Overview of my thesis

Values in Technology (Flanagan et al. 2010)

Technical inquiry

(not part of my thesis)

Empirical inquiry:

Evaluation of measurements using measurement modeling

Philosophical inquiry:

Normative discussion about bias measurements & mitigation in language models

Why measurement modeling?

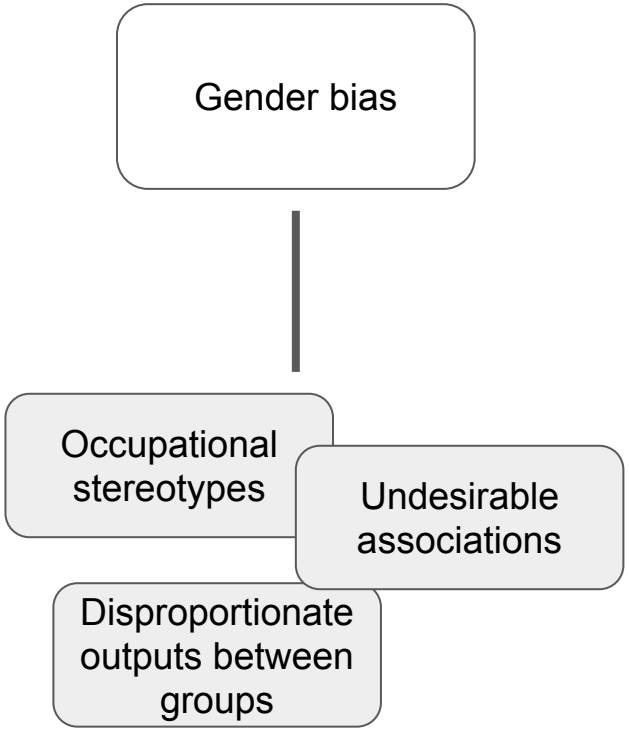
My initial observation in gender bias in language model:

1. Gender bias is mostly operationalized with occupational stereotypes
    a. Which do not necessarily capture all the relevant aspects of gender bias
2. The terminology 'bias' is used without conceptualization

Therefore, I want to critically review operationalization of bias in LM

- What (Construct) + How (Operationalization) papers are measuring gender bias?
- How to evaluate the quality of bias metrics?
    - How scientifically valid, comprehensive, and comparable these metrics are?

Evaluating the alignment



1. Construct / Conceptualization

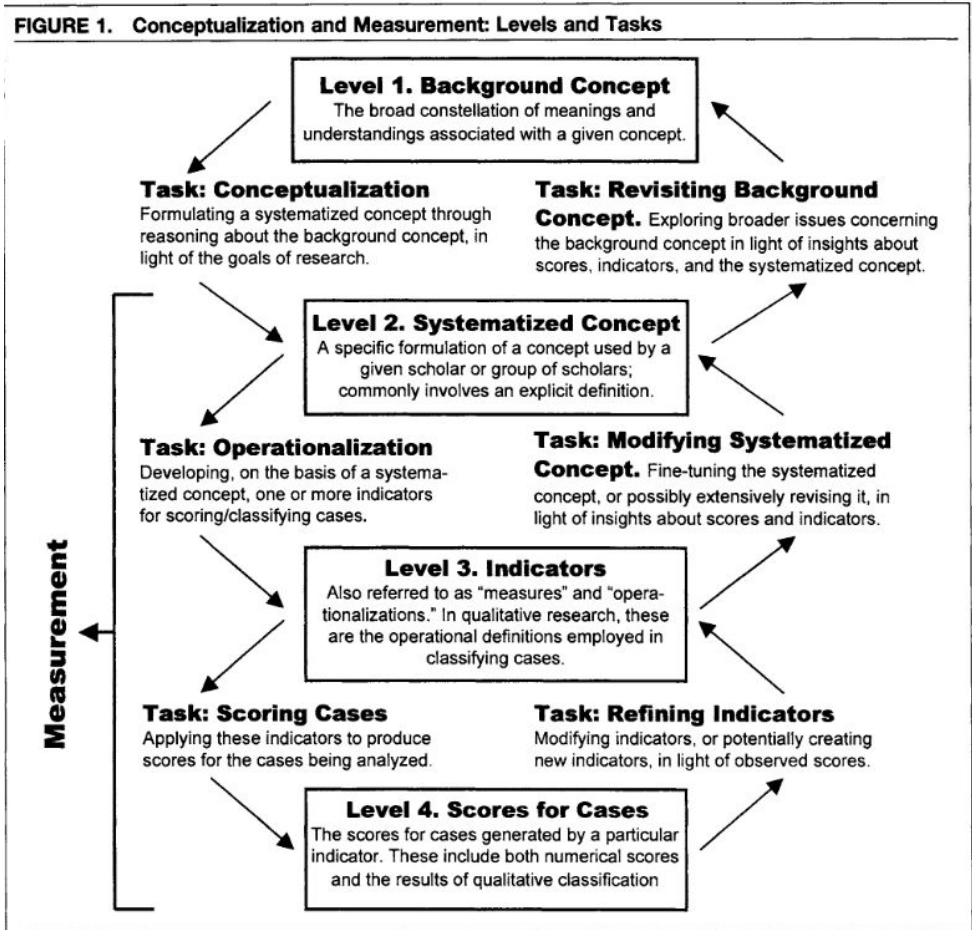
2. Operationalization / Measurement

1. Construct / Conceptualization

Measurement Validity

(Adcock and Collier, 2001, p. 529)

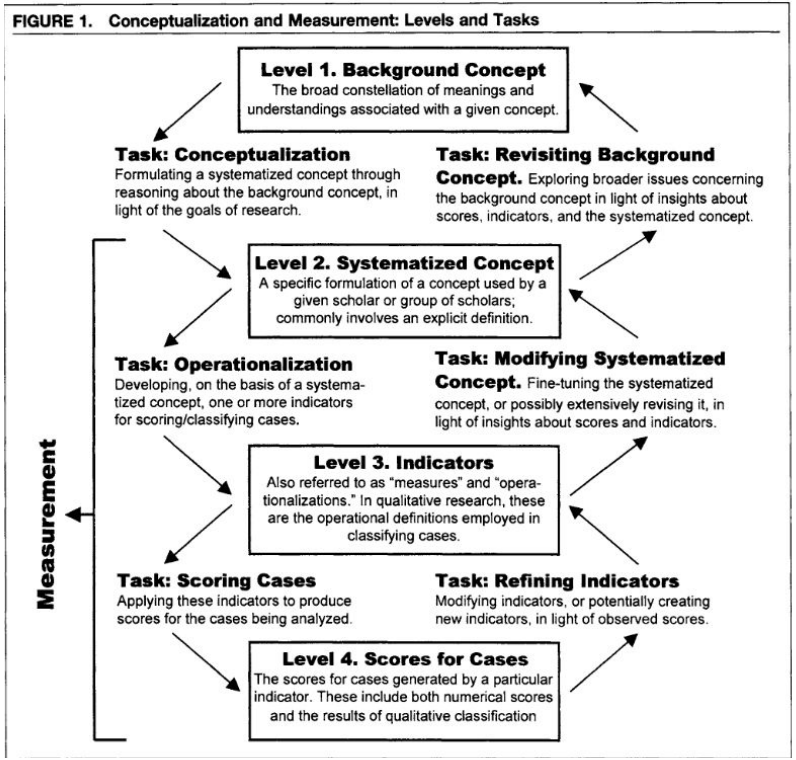
- Suggests a framework to assess more effectively, and communicate about, issues of valid measurement.
- Underscore the need to draw a clear distinction between measurement issues and disputes about concepts.
- Discusses the contextual specificity of measurement claims, exploring a variety of measurement strategies that seek to combine generality and validity by devoting greater attention to context.



Adcock and Collier (2001) p. 531

Conceptualization of gender bias in LM

- Underspecified Concept: Some works skip '`Systematized Concept`' in conceptualizing bias
- Instead, they move directly from `Background Concept` to `Indicators`
  - Many bias measurements lack formulating a systematized concept through reasoning about the background concept, in the light of goals of research



Adcock and Collier (2001) p. 531

Examples: Silva et al., 2019

Towards a Comprehensive Understanding and Accurate Evaluation of Societal Biases in Pre-Trained Transformers

Level 1: Background concepts	(Gender) Bias
Level 2: Systematized concepts	"In the context of our work, "bias" refers specifically to the preference of a model for one gender or race in the presence of an otherwise neutral context." (Silva et al., 2021, p. 1)
Level 3: Indicators	• **WEAT** (Word Embedding Association Test): "estimate implicit biases in word embeddings by measuring average cosine similarities of target and attribute sets." (Silva et al., 2021, p. 2) • **Sequence Ranking (SEQ)**: Equity Evaluation Corpus (EEC), which includes templated sequences such as "⟨TARGET⟩ feels ⟨ATTRIBUTE⟩," where gendered or racial tokens are the "targets" and emotional words are the "attributes." The average of the difference in likelihoods for target sets constitutes the bias score. • **Pronoun Ranking (PN)**: "comparing relative likelihoods of target words." (Silva et al., 2021, p. 2)

Examples: Lucy and Bamman., 2021
Gender and Representation Bias in GPT-3 Generated Stories

Level 1: Background concepts	**Gender bias (gender stereotypes, used interchangeably)** “Our work focuses on representational harms in generated narratives, especially the reproduction of gender stereotypes found in film, television, and books.” (Lucy and Bamman, 2021, p. 48)
Level 2: Systematized concepts	“We focus on **overall content differences between stories** containing prompt characters of different genders” (Lucy and Bamman, 2021, p. 50) “Even so, **depictions of women** still foreground their physical appearances (Hoyle et al., 2019), and portray them as weak and less powerful (Fast et al., 2016b; Sap et al., 2017).” (Lucy and Bamman, 2021, p. 51)
Level 3: Indicators	**Topic modeling**: “We train Latent Dirichlet allocation (LDA) on unigrams and bigrams from book excerpts and generated stories using MALLET, with 50 topics and default parameters” (Lucy and Bamman, 2021, p. 50) **Lexicon analysis:** using cosine (semantic) similarity
Level 4: Scores for cases	TM: “Feminine characters are more likely to be discussed in topics related to family, emotions, and body parts, while masculine ones are more aligned to politics, war, sports, and crime.” (Lucy and Bamman, 2021, p. 50)

Examples: May et al., 2019 (i)

On Measuring Social Biases in Sentence Encoders

Level 1: Background concepts	**Gender bias** "However, prominent word embeddings such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) **encode systematic biases against women and black people** (Bolukbasi et al., 2016; Garg et al., 2018, i.a.), implicating many NLP systems in scaling up social injustice. **We investigate whether sentence encoders, which extend the word embedding approach to sentences, are similarly biased**" (May et al., 2019, p. 622) ● **ABW stereotype** ("In the Sapphire or angry black woman (ABW) stereotype, black women are portrayed as loud, angry, and imposing (Collins, 2004; Madison, 2009; HarrisPerry, 2011; hooks, 2015; Gillespie, 2016)." ([May et al., 2019, p. 624), ● **Double Binds** ("Women face many double binds, contradictory or unsatisfiable expectations of femininity and masculinity (Stone and Lovejoy, 2004; Harris-Perry, 2011; Mitchell, 2012))
Level 2: Systematized concepts	*"A specific formulation of a concept used by a given scholar or group of scholars; commonly involves an explicit definition"* (Adcock and Collier (2001) p. 531)
Level 3: Indicators (Measures, Operationalizations)	**SEAT** (Sentence Encoder Association Test): Discrepancy in cosine similarity of vector representation of sentence ● ABW: "To measure sentence encoders' reproduction of the angry black woman stereotype, we create a test whose target concepts are **black-identifying and white-identifying female given names** from Sweeney (2013, Table 1) and whose **attributes are adjectives used in the discussion of the stereotype** in Collins (2004, pp. 87-90) and their antonyms." (May et al., 2019, p. 3) ● DB: "We test this double bind in sentence encoders by **translating Heilman et al.'s experiment** to two SEAT tests. In the first, we **represent the two target concepts by names of women and men**, respectively, in the single sentence template **"<word> is an engineer with superior technical skills."**; **the attributes are likable and non-hostile terms**, based on Heilman et al.'s design, in the sentence template **"The engineer is <word>."**" (May et al., 2019, p. 3)

Examples: May et al., 2019 (ii)
On Measuring Social Biases in Sentence Encoders

Level 1: Background concepts	**Gender bias** "However, prominent word embeddings such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) **encode systematic biases against women and black people** (Bolukbasi et al., 2016; Garg et al., 2018, i.a.), implicating many NLP systems in scaling up social injustice. **We investigate whether sentence encoders, which extend the word embedding approach to sentences, are similarly biased**" (May et al., 2019, p. 622)
Level 2: Systematized concepts	• **ABW stereotype** ("In the Sapphire or angry black woman (ABW) stereotype, black women are portrayed as loud, angry, and imposing (Collins, 2004; Madison, 2009; HarrisPerry, 2011; hooks, 2015; Gillespie, 2016)." ([May et al., 2019, p. 624), • **Double Binds** ("Women face many double binds, contradictory or unsatisfiable expectations of femininity and masculinity (Stone and Lovejoy, 2004; Harris-Perry, 2011; Mitchell, 2012))
Level 3: Indicators (Measures, Operationalizations)	**SEAT** (Sentence Encoder Association Test): Discrepancy in cosine similarity of vector representation of sentence • ABW: "To measure sentence encoders' reproduction of the angry black woman stereotype, we create a test whose target concepts are **black-identifying and white-identifying female given names** from Sweeney (2013, Table 1) and whose **attributes are adjectives used in the discussion of the stereotype** in Collins (2004, pp. 87-90) and their antonyms." (May et al., 2019, p. 3) • DB: "We test this double bind in sentence encoders by **translating Heilman et al.'s experiment** to two SEAT tests. In the first, we **represent the two target concepts by names of women and men**, respectively, in the single sentence template **"<word> is an engineer with superior technical skills."**; **the attributes are likable and non-hostile terms**, based on Heilman et al.'s design, in the sentence template **"The engineer is <word>."**" (May et al., 2019, p. 3)

Examples: Nangia et al.,

CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models

Level 1: Background concepts	
Level 2: Systematized concepts	•
Level 3: Indicators (Measures, Operationalizations)	•

2. Operationalization / Measurement

Measurement modeling (Jacobs and Wallach, 2021)

Fairness-oriented conceptualization of construct validity

Construct reliability

- Test-retest reliability

Construct validity

- Face validity
- Content validity
- Convergent validity
- Divergent validity
- Predictive / Hypothesis validity
- Consequential validity

Measurement validation (Adcock and Collier, 2001, p. 529)

Content Validation

- In the framework of Figure 1, does a given indicator (level 3) adequately capture the full content of the systematized concept (level 2)?
- First, are key elements omitted from the indicator? Second, are inappropriate elements included in the indicator? An examination of the scores (level 4) of specific cases may help answer these questions about the fit between levels 2 and 3.

Convergent/Discriminant validation

- Are the scores (level 4) produced by alternative indicators (level 3) of a given systematized concept (level 2) empirically associated and thus convergent?

Nomological/Construct validation

- In a domain of research in which a given causal hypothesis is reasonably well established, we ask: Is this hypothesis again confirmed when the cases are scored (level 4) with the proposed indicator (level 3) for a systematized concept (level 2) that is one of the variables in the hypothesis?

Measurement modeling (Jacobs and Wallach, 2021)
Fairness-oriented conceptualization of construct reliability

Construct reliability: “do similar inputs to a measurement model, possibly presented at different points in time, yield similar outputs?” (Jacobs and Wallach, 2021, p. 378)

1. Test-retest reliability: “the extent to which measurements of an unobservable theoretical construct, obtained from a measurement model at different points in time, remain the same, assuming that the construct has not changed.” (Jacobs and Wallach, 2021, p. 379)

Measurement modeling (Jacobs and Wallach, 2021)
Fairness-oriented conceptualization of construct validity

Construct validity: “demonstrating, in a variety of ways, that the measurements obtained from measurement model are both meaningful and useful” (Jacobs and Wallach, 2021, p. 379-381)

“the extent to which the measurements obtained from a measurement model…”

1. Face validity:
   - “look plausible - a “sniff test” of sorts.”
2. Content validity:
   - “an operationalization wholly and fully captures the substantive nature of the construct purported to be measured.”
3. Convergent validity:
   - “…correlate with other measurements of the same construct, obtained from measurement models for which construct validity has already been established.”
4. Discriminant validity:
   - “… vary in ways that suggest that the operationalization may be inadvertently capturing aspects of other constructs.”
5. Predictive validity:
   - “… are predictive of measurements of any relevant observable properties (and other unobservable theoretical constructs) thought to be related to the construct purported to be measured, but not incorporated into the operationalization.”
6. Hypothesis validity:
   - “… support substantively interesting hypotheses about the construct purported to be measured.”
7. Consequential validity:
   - “concerned with identifying and evaluating the consequences of using the measurements obtained from a measurement model, including any societal impacts.”

Questions for testing assumptions (Jacobs and Wallach 2021)

1. Face validity: Do the measurements look plausible?
2. Content validity
    1. Does the operationalization capture all relevant aspects of the construct purported to be measured?
    2. wholly and fully captures the substantive nature of the construct purported to be measured?
        1. contestedness: does it have multiple context-dependent, conflicting, theoretical understandings? - inherently hard to assess content validity; **unlikely that a single operationalization can wholly and fully capture its substantive nature in a meaningful fashion.**
        2. substantive validity
            1. does the measurement modeling process - i.e., the assumptions made when moving from abstractions to mathematics incorporate measurements of those - and only those - observable properties (and other unobservable theoretical construct, if appropriate) thought to be related to the construct?
        3. structural validity
            1. operationalization captures the structure of the relationships between the incorporated observable properties (and other unobservable theoretical constructs, if appropriate) and the construct purported to be measured, as well as the interrelationships between them?
3. Convergent validity
    1. Do they correlate with other measurements of the same construct, obtained from measurement models for which construct validity has already been established?
    2. danger of yielding a false sense of security when correlating with measurements that have not been sufficiently well validated
4. Discriminant validity
    1. measurements obtained from a measurement model vary in ways that suggest that the operationalization may be inadvertently capturing aspects of other constructs.
5. Predictive validity
    1. Are the measurements predictive of measurements of any relevant observable properties (and other unobservable theoretical constructs) thought to be related to the construct, but not incorporated into the operationalization?
    2. concerned with the utility of the measurements, not their meaning
6. Hypothesis validity
    1. support substantively interesting hypotheses about the construct purported to be measured?
    2. concerned with the utility of the measurements, not their meaning
    3. not always clear cut from predictive validity
7. Consequential validity
    1. What are the consequences of using the measurements - including any societal impacts.
    2. identifying and evaluating the consequences of using the measurements obtained from a measurement model, including any societal impacts.
    3. “measurements both reflect structure in the natural world, and impose structure upon it.”
    4. How is the world shaped by using the measurements? What would do we wish to live in? If there are contexts in which the consequences of using the measurements would cause us to compromise values that we wish to uphold, then the measurements should not be used in those contexts.

Intro to ACSS slides

Validity and Reliability

- (Construct) Reliability: consistency of a measure.
  - If we were to redo our measurement now, or again tomorrow, or on a similar test set, would we recover similar results?
- (Construct) Validity: does a method measures what it is intended to measure?
  - Does the measurements behave as expected? (Face Validity)
  - Does the measurement capture all relevant facets of the construct? (Content Validity)
  - Does the measurement match other accepted measurements of this construct or other external criteria? (Criterion Validity and Convergent Validity)
  - Do our measurements match other related constructs? (Discriminant Validity)

Example: Sexism

- Face Validity: take extreme examples (e.g. #feminismIsCancer)
- Content Validity: use examples from survey scales
- Convergent Validity: compare our measurement with other sexism measurement methods
- Discriminant Validity: is our measurement also measuring related constructs (e.g. hate)?
- Reliability: do we get similar results if we repeat our measurement, e.g. on different datasets?

Applying measurement modeling to bias metrics

Construct reliability

- Test-retest reliability: Not tested; some provide statistical significance score like correlation coefficient, t-test, etc.

Construct validity

- Face validity: Most of them plausible - not completely unrelated or random
- Content validity: Some uses multiple metrics (reflecting multifaceted nature of the concept)
- Convergent validity, Divergent validity:
    - Some compares with other benchmarks
    - Otherwise difficult to evaluate due to the absence of established measurements of the construct in LM;
- Predictive / Hypothesis validity: difficult to apply
- Consequential validity: Not discussed; most measurements do not consider the risk of using specific metrics to measure gender bias
    - Some note that measurements confirm the presence of bias, but does not guarantee lack of bias when it is not ‘measured’ with the metric

Categories of bias measurements in LM

- **External scoring**
  - Toxicity (Dhamala et al., 2021)
  - Sentiment (Dhamala et al., 2021, Jentzsch and Turan, 2022)
  - Psycholinguistic norms (Dhamala et al., 2021)
- **Embedding analysis**
  - May et al., 2019, Silva et al., 2021, Wolfe and Caliskan, 2021
- **Probability of masked tokens**
  - Stereotypical, anti-, unrelated (Nangia et al., 2020)
  - Pair of sentences differing with monosex names: (Steinborn et al., 2022)
  - Kaneko et al., 2022, Touileb et al., 2022, Nadeem et al., 2019, Silva et al., 2021, Alnegheimish et al., 2022, de Vassimon Manela et al., 2021
- **Performance in downstream tasks**
  - Coreference resolution
  - Classification
  - NLI, CI (Sotnikova et al., 2021)
  - Sequence ranking (Silva et al., 2021)
  - Pronoun ranking (Silva et al., 2021)
  - Price recommendation (Shen et al., 2023)
  - Dialogue state tracking Barikeri et al., 2021
  - Conversational response-generation (Barikeri et al., 2021
  - Perplexity-based score (Barikeri et al., 2021)
- **Semantic analysis**
  - Analyze generated text
    - Topic modeling, Lexical analysis: Lucy and Bamman 2021

Nangia et al., 2020 (CrowS-pairs): **Stereotype pseudo-log-likelihood of masked token (ste. vs. anti-ste. vs. unrelated)**

Construct reliability

- Test-retest reliability: Not tested

Construct validity

- Face validity: Plausible
- Content validity: Stereotype does not capture all relevant aspect of social bias
- Convergent validity, Divergent validity: Compare with `WinoBias` and `StereoSet` as baselines, and found that all three models exhibit substantial bias.
- Predictive / Hypothesis validity: Model's preference towards stereotypical sentences might be related to other observables such as translating or generating sentences with occupational bias…?
- Consequential validity: Risk of equating social(gender) bias into stereotypes

Lucy and Bamman, 2021: **Gender and Representation Bias**

Topic modeling and Lexical analysis

Cooccurence with topical terms **(e.g. family, sports, politics, body parts, etc.)**

Construct reliability

- Test-retest reliability: Not tested, but statistical significance was tested (Pearson r, t-test with Bonferroni correlation, $p < 0.05$)

Construct validity

- Face validity: Plausible
- Content validity: Builds upon Sap et al., (2017) on power of characters in film
- Convergent validity, Divergent validity: Use two different method that show similar result
- Predictive / Hypothesis validity: Probably other measurements of gender bias will align?
- Consequential validity: Semantic analysis - societal impact would be smaller than other metric-based method

May et al. 2019: **ABW stereotype, Double-bind**
Discrepancy between cosine similarities

Construct reliability

- Test-retest reliability: Not tested; but effect size, p-value were provided

Construct validity

- Face validity: Plausible; undesirable stereotypes relate to gender bias
- Content validity: yes; the authors aim to measure ABW, double-bind stereotypes in LM; based on established literature on those biases (Sweeny 2013, Collins 2004, Heilman et al., 2014)
- Convergent validity: Divergent validity: what would be other established measurement of ABW, DB? Maybe the existence of ABW in LM was validated by other studies?
- Predictive / Hypothesis validity: what would be other relevant observable properties/unobservable theoretical construct of ABW, DB? Should I refer to the original theoretical works the authors refer to in designing the measurements for ABW and DB?
- Consequential validity: Relies on theoretical work the authors refer to as a theoretical basis?

Silva et al. 2019: **ABW stereotype, Double-bind**
Cosine similarity, Embedding

Construct reliability

- Test-retest reliability: Not tested; but effect size, p-value were provided

Construct validity

- Face validity: Plausible
- Content validity: Stereotype does not capture all relevant aspect of social bias
- Convergent validity:
- Divergent validity: Compare with `WinoBias` and `StereoSet` as baselines, and found that all three models exhibit substantial bias. (any correlation?)
- Predictive / Hypothesis validity: Model's preference towards stereotypical sentences might be related to other observables such as translating or generating sentences with occupational bias…?
- Consequential validity: Risk of equating social(gender) bias into stereotypes

Dhamala et al., 2021:

Biases related to occupational associations for different protected categories

Sentiment, Toxicity, Gender Polarity, Regard, Psycholinguistic norms

Construct reliability

- Test-retest reliability: Not tested

Construct validity

- Face validity: Plausible
- Content validity: Stereotype does not capture all relevant aspect of social bias
- Convergent validity:
- Divergent validity: Compare with `WinoBias` and `StereoSet` as baselines, and found that all three models exhibit substantial bias. (any correlation?)
- Predictive / Hypothesis validity: Model's preference towards stereotypical sentences might be related to other observables such as translating or generating sentences with occupational bias…?
- Consequential validity: Risk of equating social(gender) bias into stereotypes

Challenges / Limitations

In order to fit into the framework, I need to infer what is not explicitly in the paper

How to integrate the table into the thesis?

- Include the table + focus on the most relevant part according to the topic in text

  Full paper list

Relevant work: Applying Measurement Modeling to NLP

Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets (Blodgett et al., 2021)

- Apply measurement modeling to 4 benchmark datasets
  - CrowS-Pairs, StereoSet, WinoBias, WinoGender
- Show limitations on conceptualization and operationalization through lack of clear articulations of construct and a range of ambiguities and unstated assumptions

Trustworthy Social Bias Measurement (Bommasani and Liang, 2022)

- Operationalize bias by proposing a general bias measurement framework *DivDist*,
- Propose a testing protocol with 8 testing criteria (e.g. predictive validity: do measures predict biases in US employment?)