The Fill-Mask Association Test (FMAT): Measuring Propositions in Natural Language

Author: Han-Wu-Shuang Bao

Affiliations:

[1] School of Psychology and Cognitive Science, East China Normal University

[2] Manchester China Institute, University of Manchester

[*] Correspondence to: Han-Wu-Shuang Bao (baohws@foxmail.com)

Citation: Bao, H.-W.-S. (in press). The Fill-Mask Association Test (FMAT): Measuring propositions in natural language. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspa0000396

Notes: All data, analysis code, and supplemental results are available at https://osf.io/5e2hr/. Details about the R package "FMAT" can be found at https://psychbruce.github.io/FMAT/.

Acknowledgments: I gratefully acknowledge the support and resources provided by the Manchester China Institute. I thank these scholars for their helpful feedback and comments: Huajian Cai, Peter Gries, Todd Hartman, Tingting Huang, Meijia Li, Li Lin, Bo Wang, Xiaobing Wang, Zi-Xi Wang, Arkadiusz Wiśniowski, Samuel Ying Yang, Xiaolin Zhou, and the editor and reviewers. I also dedicate this article to the memory of Keita Kurita, who followed a parallel intellectual path but passed away in 2020.

Correspondence: Han-Wu-Shuang Bao, School of Psychology and Cognitive Science, East China Normal University, 3663 North Zhongshan Road, Shanghai 200062, China, and Manchester China Institute, University of Manchester, 178 Waterloo Place, Oxford Road, Manchester M13 9PL, United Kingdom. Email: baohws@foxmail.com

(Title page updated April 12, 2024 to correct the automatic identification by Google Scholar)

The Fill-Mask Association Test (FMAT): Measuring Propositions in Natural Language

Han-Wu-Shuang Bao

Abstract

Recent advances in large language models are enabling the computational intelligent analysis of psychology in natural language. Here, the Fill-Mask Association Test (FMAT) is introduced as a novel and integrative method leveraging Masked Language Models to study and measure psychology from a *propositional* perspective at the societal level. The FMAT uses BERT models to compute semantic probabilities of option words filling in the masked blank of a designed query (i.e., a cloze-like contextualized sentence). The current research presents 15 studies that establish the reliability and validity of the FMAT in predicting factual associations (Studies 1A–1C), measuring attitudes/biases (Studies 2A–2D), capturing social stereotypes (Studies 3A–3D), and retrospectively delineating lay perceptions of sociocultural changes over time (Studies 4A–4D). Empirically, the FMAT replicated seminal findings previously obtained with human participants (e.g., the Implicit Association Test) and other big-data text-analytic methods (e.g., word frequency analysis, the Word Embedding Association Test), demonstrating robustness across 12 BERT model variants and diverse training text corpora. Theoretically, the current findings substantiate the propositional (vs. associative) perspective on how semantic associations are represented in natural language. Methodologically, the FMAT allows for more fine-grained language-based psychological measurement, with an R package developed to streamline its workflow for use on broader research questions.

Keywords: natural language processing, large language models, propositional representation, attitudes, social cognition

> "There is nothing so practical as a good theory."
>
> —*Kurt Lewin* (1951)
>
> "There is nothing so theoretical as a good method."
>
> —*Anthony G. Greenwald* (2012)

With the rapid growth of artificial intelligence (AI), we are experiencing a surge of interest in large language models (LLMs) that can understand and generate human-like language, and in how they can facilitate social science research (e.g., Argyle et al., 2023; Cutler & Condon, 2023; Dillion et al., 2023; Grossmann et al., 2023; Wang et al., 2023). Most of modern LLMs have evolved from two mainstream language models: GPT and BERT (Yang et al., 2023). OpenAI's GPT (Generative Pre-trained Transformer) is trained to generate text based solely on the antecedent words using an autoregressive, unidirectional, and open-ended approach (Radford et al., 2019). In contrast, Google's BERT (Bidirectional Encoder Representations from Transformers) is trained to predict masked words in a sentence while considering both the left and right contexts, allowing the model to develop a deeper understanding of the relationships between words and the contexts where they are used (Devlin et al., 2018).

LLMs like GPT and BERT are not search engines that simply count words in texts, but are trained with deep learning to understand any new contexts and provide *semantically probable* responses (Berger & Packard, 2022; Rogers et al., 2020; Yang et al., 2023). LLMs contain large parameters representing human knowledge, thoughts, and feelings inherited from the texts on which the models are trained. Thus, LLMs can provide semantic (rather than "realistic") summaries of what a given group of people wrote in a specific corpus of texts at a particular time. Such semantic responses can therefore be understood as how an average person in that population would respond to specific queries, enabling the study of human psychology in natural language without recruiting human participants (Dillion et al., 2023; Grossmann et al., 2023).

However, *how* LLMs can be used to better understand human psychology, society, and culture remains a challenge. Leveraging BERT models and adopting the propositional

perspective on attitudes and social cognition (De Houwer et al., 2020, 2021), the current research introduces a novel and integrative method: the *Fill-Mask Association Test* (FMAT). Here, the term "association" is used in an inclusive sense, referring to conceptual relations or associations, but especially with specific relational information (see Research Questions for the different theoretical perspectives). A series of 15 studies were conducted to evaluate the FMAT method's psychometric properties and test whether it can replicate a variety of seminal findings previously obtained with human participants and/or other text-analytic methods.

To contextualize the FMAT within the progress of language analysis in psychology, advances in natural language analysis methods are first reviewed from a measurement perspective, identifying the key methodological limitations of existing approaches. Then, the FMAT is introduced to show how it can address these limitations and extend the propositional perspective to the natural language study of attitudes and social cognition. In doing so, the current research makes both methodological and theoretical contributions. Furthermore, the FMAT should travel well to new questions and broader areas of research.

Language as Measurement

Psychological Measurement: From the Individual to Societal Levels

Human psychology can be measured quantitatively at both individual and societal levels. To assess individual differences, decades of research have used two main approaches: (1) *direct* self-report measures, such as Likert scales (Likert, 1932) and semantic differential scales (Osgood et al., 1957), which can assess explicit thoughts, beliefs, and emotions; and (2) *indirect* measures, such as the Implicit Association Test (IAT; Greenwald et al., 1998) and its diverse variants, which aim to tap into implicit psychological processes. For instance, the IAT requires participants to complete a computer-based key-response task to categorize a set of target concepts and attribute words. The difference in response latency or categorization errors between compatible and incompatible conditions can indicate how strongly a person associates a target concept with an attribute dimension in their mind (Greenwald et al., 1998). The reliability of such implicit measurement ranges from high to low (for reviews, see Fazio & Olson, 2003; Gawronski & De Houwer, 2014; Gawronski et al., 2020; Nosek et al., 2011).

The above methods aim to measure the psychology *inside* an individual's head; it is also essential and feasible to measure psychology *outside* the head at a broader societal level. To this end, a promising and widely adopted approach is to analyze language, particularly human-generated texts recorded in tangible and public cultural products (Jackson et al., 2022; Morling & Lamoreaux, 2008). Language/text analyses allow for more objective observation of people's natural expressions of their thoughts and feelings, thereby measuring psychology with less response bias and greater efficiency (Berger & Packard, 2022; Grossmann et al., 2023; Jackson et al., 2022).

Using Natural Language to Study Social Psychology

Natural language, the ways people naturally talk and write in the real world, conveys rich information about what and how people think and feel about each other. Recent advances in natural language processing (NLP) enable us to analyze people's discourse quantitatively and more objectively. Quantitative methods of natural language analysis can be broadly classified into three distinct approaches: word counting, word embedding, and language modeling.

The Word-Counting Approach

Language analysis, at first, involved simply counting the frequencies of a preselected list of words (namely, a "dictionary"). The basic idea behind this method is that language use can reflect individual differences and sociocultural characteristics (Pennebaker & King, 1999; Pennebaker et al., 2003). Over decades, this assumption has gained wide acceptance, with extensive studies stimulated by dictionary-based tools for language analysis, such as the Linguistic Inquiry and Word Count (LIWC; Tausczik & Pennebaker, 2010), and large-scale digitized text corpora, such as the Google Books Ngram (Michel et al., 2011). By employing these tools and databases, together with custom dictionaries created by individual researchers, studies have yielded meaningful findings. For example, word frequency analyses have found an increase in the use of words reflecting individualism versus collectivism over the past two centuries (e.g., Greenfield, 2013; Grossmann & Varnum, 2015), consistent with findings from self-report measures and societal indicators such as family structure (Santos et al., 2017).

The word-counting approach is simple and fast, allowing for flexible analyses of word use from small collections of texts to larger corpora (Pennebaker et al., 2003; Tausczik & Pennebaker, 2010). However, word frequency indicates only the prevalence or popularity of a concept but not people's endorsement or acceptance, precluding it from being used to address deeper theoretical questions. Furthermore, word counting has little access to semantic and contextual information, making it unlikely to analyze semantic relatedness or clarify what meanings people *intend* to express through word use. In addition, it is vulnerable to selection bias arising either from biased inclusion criteria of text corpora (e.g., present in the Google Books database; Varnum & Grossmann, 2017) or from "researcher degrees of freedom" in selecting (cherry-picking) certain words to arrive at more favorable results (Simmons et al., 2011). Several recent articles discuss more thoroughly the limitations and challenges faced by the word-counting approach (e.g., Atari & Henrich, 2023; Berger & Packard, 2022; Boyd & Schwartz, 2021; Jackson et al., 2022).

The Word-Embedding Approach

To enable machine understanding of human language, a basic strategy is to quantify the meaning of a word through nearby words that often accompany it (Harris, 1954). As the *distributional semantic* hypothesis posits, words used in similar contexts have more similar semantic meanings (Lenci, 2018). Earlier methods for distributional semantics involve topic modeling, with either Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) or Latent Dirichlet Allocation (LDA; Blei et al., 2003) as a statistical way to reduce the dimensionality of a word-by-document co-occurrence matrix (Lenci, 2018). Hence, some latent categorical structures (e.g., semantic dimensions or topics) can be extracted from texts, with a word's meaning represented by a low-dimensional numeric vector (Berger & Packard, 2022).

Recent advances rely more on deep learning (e.g., neural networks) to "embed" word semantics in a continuous vector space, translating words into numeric vectors (termed word embeddings) that quantify their semantic meanings (Bengio et al., 2003). For example, the Word2Vec algorithm converts words into vectors by predicting words based on surrounding words, or vice versa (Mikolov et al., 2013); the GloVe algorithm produces word vectors by

predicting word co-occurrences in a whole corpus (Pennington et al., 2014). The cosine of the angle between two word vectors quantifies *semantic similarity*, indicating how the two words are used in similar contexts (Word2Vec) or co-occur in a corpus of texts (GloVe).

Word embeddings have recently been used in disciplines throughout and well beyond psychology. A landmark publication in *Science* introduced the Word Embedding Association Test (WEAT) as a method using semantic similarity between word embeddings to measure human-like biases in natural language (Caliskan et al., 2017). Using the WEAT, Caliskan et al. (2017) replicated a spectrum of classic findings in social psychology originally obtained with the IAT (Greenwald et al., 1998), including attitudes (e.g., toward flowers vs. insects), social biases (e.g., toward European vs. African Americans), and stereotypes (e.g., the gender stereotype associating men with career and women with family); furthermore, they captured factual associations that can predict real gender distributions of occupations and first names. Since then, a rapidly growing number of studies have used the WEAT to assess social biases and stereotypes (e.g., Bailey et al., 2022; Charlesworth et al., 2022; DeFranza et al., 2020; Napp, 2023), and even to track changes in stereotypes (e.g., Garg et al., 2018) and cultural-psychological associations (e.g., Bao et al., 2022) with decade-specific word embeddings.

Semantic similarity analyses of word embeddings have yielded rich insights for social and cultural psychology. However, recent studies have raised concerns about their validity and reliability. A major critique is that word embeddings can hardly capture the goals, desires, and beliefs that people express through words (Lake & Murphy, 2023). For example, a high semantic similarity between “I” and “happy” is difficult, if not impossible, to distinguish “I AM happy” (an actual affect) from “I WANT TO BE happy” (an ideal affect). Indeed, *static* word embeddings such as Word2Vec and GloVe cannot address any contextual information or disambiguate words with multiple meanings (Sabbaghi et al., 2023). Another concern with word embeddings is the frequency-based distortion: more frequent words can produce higher semantic similarity of word embeddings, even with a constant distribution of co-occurrences (Valentini et al., 2023); and word embeddings tend to cluster frequent (vs. rare) words with positive (vs. negative) words, producing spuriously more positive bias toward more frequent

terms (van Loon et al., 2022). Accordingly, the WEAT may systematically overestimate the magnitude of bias due to word frequency distortion (exaggeration), sometimes even due to arbitrary word selection (Ethayarajh et al., 2019). In addition, a single bias measure derived from word embeddings may suffer from low internal consistency between words (Silva et al., 2021) and poor inter-rater agreement among scoring rules (Du et al., 2021).

The Language-Modeling Approach

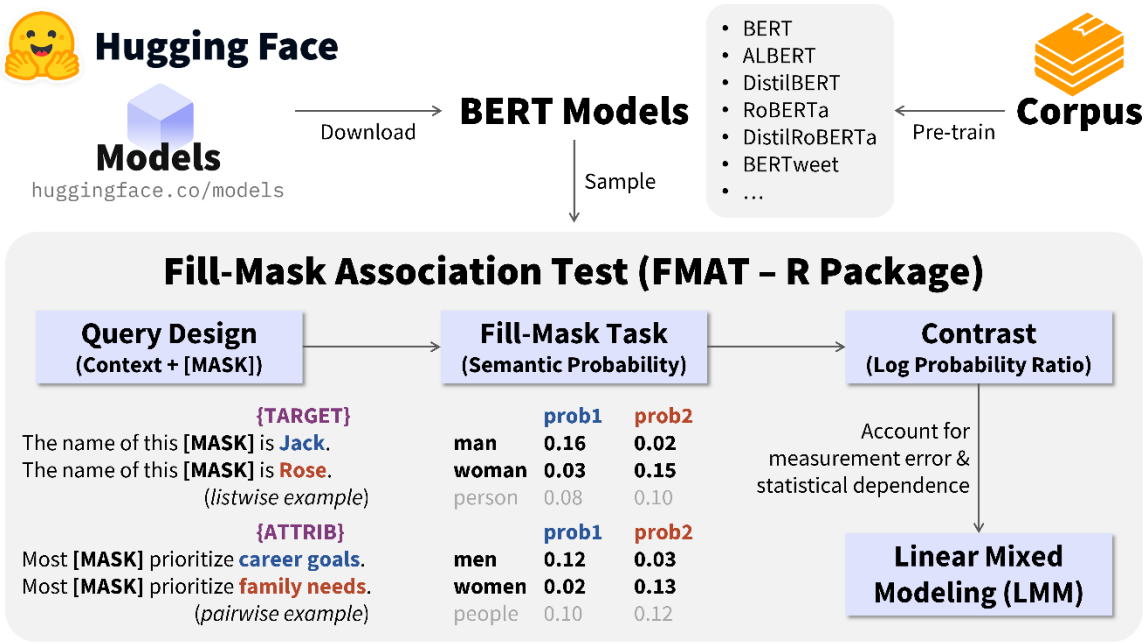
The method introduced in the current research—FMAT—adopts a new text-analytic approach: language modeling (Berger & Packard, 2022; Dillion et al., 2023), which differs from word embedding in that language models can directly process contextual information. The FMAT utilizes BERT models, a family of language understanding models built with Google's bidirectional Transformer architecture (Devlin et al., 2018). Inspired by the *cloze* task (Taylor, 1953), a BERT model is pre-trained with Masked Language Modeling to predict masked words in a sentence given the context specified. Pre-trained BERT models inherit semantic, syntactic, and world knowledge, with an ability to capture semantic and relational information in new contexts (Rogers et al., 2020; Wang et al., 2023; Yang et al., 2023). For example, given the query "Paris is the [MASK] of France.", a BERT model can predict the most likely answer—"capital"—with an estimated *semantic probability* (i.e., the conditional probability of a word filling in the mask given the context). Beyond linguistic knowledge, BERT models also inherit human psychological, social, and cultural information from the training text corpora, reflecting constructs such as personality structure (Cutler & Condon, 2023), moral norms (Schramowski et al., 2022), and social biases (Bartl et al., 2020; Kaneko & Bollegala, 2022; Kurita et al., 2019; May et al., 2019; Nadeem et al., 2020; Nangia et al., 2020; Silva et al., 2021). More importantly, BERT models allow for designing naturalistic queries that can specify relational information and thus more closely mirror the form and spirit of survey, better than context-free word analyses for studying psychology in language (Argyle et al., 2023; Berger & Packard, 2022; Cutler & Condon, 2023; Dillion et al., 2023; Grossmann et al., 2023; Wang et al., 2023; Widmann & Wich, 2023).

As a novel foray into the language-modeling approach, the FMAT uses BERT models

to estimate the semantic probabilities of option words filling in the mask of a query, which are then contrasted between conditions to test relative conceptual associations in language (see Figure 1 for the FMAT workflow). Generally, contrasts (either pairwise or listwise) need to be designed for *both* query contextual phrases and masked words to partial out confounds (e.g., superordinate concepts, disproportionate word frequencies; see Methodology Overview Step 2 for details). On the one hand, the query contexts are comparable phrases, with targets or attributes labeled as {TARGET} or {ATTRIB} (conceptually distinct, but technically interchangeable). On the other hand, the [MASK] options are comparable words for attributes (if contexts are {TARGET}) or targets (if contexts are {ATTRIB}). For example, to test the "Male = Career, Female = Family" gender stereotype, the query can be set as "Most [MASK] prioritize {ATTRIB}." where the {ATTRIB} is specified as "career goals" vs. "family needs" (attributes) and the [MASK] options are "men" vs. "women" (targets). Examples are endless, but the current research attempts to provide diverse instances for various purposes.

Figure 1

Workflow of the Fill-Mask Association Test (FMAT) Proposed in the Current Research



Note. The FMAT query should be a grammatically correct sentence template, with {TARGET} or {ATTRIB} (interchangeable) specifying main contextual phrases, and with [MASK] for BERT models to estimate the semantic probability (*not* actual frequency) of each optional filler word. Depending on purpose, contextual phrases and [MASK] options can be contrasted pairwise or listwise.

The Current Research

Research Questions

The current research addresses two questions. The first is methodological: How well can the FMAT measure semantic associations in natural language? The FMAT is proposed as a novel, integrative, and versatile method for measuring semantic associations that may imply psychological, social, and cultural associations. While preliminary work has explored bias in BERT models following a similar fill-mask approach (Bartl et al., 2020; Kurita et al., 2019), the current work goes beyond bias and systematically examines the reliability and validity of the new FMAT method. Reliability was evaluated using (1) the internal consistency among FMAT queries (Cronbach's α_{query}) and (2) the inter-rater agreement among BERT models (intraclass correlation coefficient, ICC), with ICC_{single} indicating the reliability of a single BERT model and $ICC_{average}$ indicating the reliability of average results across all sampled BERT models (Shrout & Fleiss, 1979).[1] Validity was appraised with (1) criterion-related validity (with gold standards such as factual information and the IAT), (2) convergent validity (with former text-analytic methods such as the WEAT), (3) discriminant validity (to capture the hypothesized constructs but not others), and (4) incremental validity (over the other methods). Furthermore, an R package "FMAT" was developed (Bao, 2023). Thus, the current research makes methodological contributions by establishing the psychometric properties of

[1] The ICC analysis treated (1) log probabilities of words filling in the mask of a query as "rating scores"; (2) *n* uniquely filled sentences as "rating items" (rows); and (3) *k* BERT models as the "raters" (columns). Below are the formulas of both types of ICCs (McGraw & Wong, 1996). Mathematically, ICC_{single} always contains a greater denominator and thus is always smaller than $ICC_{average}$ (Shrout & Fleiss, 1979, p. 426). Hence, a more empirically meaningful way is to test ICCs against certain criteria. ICCs above .60 and .75 are typically interpreted as "good" and "excellent" agreement, respectively (Cicchetti, 1994). In the current research, ICCs and their 95% confidence intervals (CIs) were estimated using the "icc()" function from the R package "irr" (Gamer et al., 2019). See online supplemental materials for details about the R code.

$$ICC_{single} = \frac{MS_{row} - MS_{error}}{MS_{row} + \frac{k}{n}(MS_{column} - MS_{error}) + (k-1)MS_{error}}$$

$$ICC_{average} = \frac{MS_{row} - MS_{error}}{MS_{row} + \frac{1}{n}(MS_{column} - MS_{error})}$$

FMAT and streamlining its workflow for easier and standardized use in future research.

Second, this research addresses a conceptual question with theoretical implications for the study of attitudes and social cognition: Are semantic associations stored in an *associative* or *propositional* way in text? More specifically, is the propositional perspective applicable to natural language? To take attitudes as an example, suppose one aims to detect a positive attitude toward Peter, which can basically be characterized as a semantic association between "Peter" and positivity (vs. negativity). From the *associative* perspective, such an attitude is simply a stronger link between the target "Peter" and a positive attribute (e.g., "pleasure") compared to a negative attribute (e.g., "disaster"), without an intention to evaluate the target (Gawronski & Bodenhausen, 2006, 2011). However, from the *propositional* perspective, this attitude is represented as a proposition in nature that specifies relational information—*how* the targets and the attributes are related to each other—and thus can also be evaluated as either true or false (De Houwer et al., 2020, 2021; Gawronski & Bodenhausen, 2006, 2011). For example, the positive attitude toward Peter can be translated into a more concrete statement such as "It's been a *pleasure* meeting *Peter*" or "We *like Peter*" rather than "*Peter* listens to music for *pleasure*" or "*Peter likes* watching *disaster* movies." Several implicit tasks have been developed to measure attitudes from the propositional perspective, based on relational responses of human participants.[2]

Applying the propositional perspective to semantic associations in natural language has theoretical and methodological implications. Theoretically, since human-generated text is essentially a collection of written statements, semantic meanings are arguably stored as propositions in texts. Indeed, a grammatically correct, semantically meaningful statement in natural language often involves a proposition, not just the co-occurrence of words without

[2] Relational responses of human participants can be measured by the Relational Responding Task (RRT; De Houwer et al., 2015), the Implicit Relational Assessment Procedure (IRAP; see Barnes-Holmes et al., 2010 for a review), the Natural Language IRAP (Kavanagh et al., 2016), the autobiographical IAT (aIAT; see Agosta & Sartori, 2013 for a review), and the questionnaire-based IAT (qIAT; Friedman et al., 2021; Yovel & Friedman, 2013). Compared to traditional IATs, these measures are conceptually more similar to the FMAT in that they all measure attitudes with *propositions* (rather than single words without relational information).

relational information. Hence, the propositional perspective originating from attitude research (De Houwer et al., 2020, 2021) can lay the foundation for a more generic and authentic view of natural language semantics, promoting the understanding of most psychological constructs that can be measured by text. Methodologically, the propositional perspective also allows for more concrete measurement of specific relations in texts and for detecting nuances between contexts. Notably, although word embeddings can quantify word co-occurrence patterns, such patterns are derived from propositions rather than simple accumulations of words—and word embeddings cannot specify *how* words are related to each other. In contrast, language models like BERT effectively encode deep relational information, enabling the study of propositions in language (Cutler & Condon, 2023; Rogers et al., 2020; Wang et al., 2023). Taken together, there is an empirical need to test (i.e., substantiate with concrete instances) the propositional perspective on semantic representations in natural language. To this end, the current research introduces and uses *propositional queries* (i.e., masked query sentences that specify relational information, e.g., "I [MASK] piano." [*like* vs. *dislike*], rather than ambiguous statements) to test associations or relations (e.g., factual associations, attitudes, biases, stereotypes) that might be previously understood as only associative links.[3]

In addition to methodological and theoretical questions, the current research also aims to demonstrate the practical value of FMAT for potential application to new lines of research. The FMAT is expected to leverage the advantages of BERT models to capture more complex semantic relationships and thus probe many advanced psychological constructs (e.g., goals, desires, beliefs, interests, social norms, intersectional stereotypes, prescriptive stereotypes).

---

[3] The associative vs. propositional perspectives should also be disentangled from the implicit vs. explicit (automatic vs. controlled) processes (Gawronski & Bodenhausen, 2009). Implicit evaluations, in which stimuli automatically elicit human evaluative responses, can also be based on the formation or activation of propositional representations (e.g., De Houwer, 2006, 2014; De Houwer et al., 2021; Moran et al., 2022). While the current research focuses on and adopts the propositional perspective, it should be acknowledged that natural language may encode both explicit and implicit attitudes (Wang et al., 2019), though evidence also shows that WEAT attitude scores and implicit (but not explicit) attitudes are positively and strongly correlated across topics (Morehouse et al., 2023).

Table 1

Overview of the 15 Studies in the Current Research

Study	Effect	[MASK] word	{TARGET} or {ATTRIB}	Example query sentence (with replaced {TARGET}/{ATTRIB} shown in bold)
Study 1	*Factual associations*			
1A	Occupation-gender association	Gender	50 occupations	"The [MASK] works as a **nurse**." [*man* vs. *women*]
1B	Name-gender association	Gender	50 names	"The name of this [MASK] is **Jackie**." [*man* vs. *women*]
1C	Name-gender association	Gender	3,644 names	"The name of this [MASK] is **Jackie**." [*man* vs. *women*]
Study 2	*Attitudes and social biases*			
2A	Flower–insect attitude	Attitude (or non-attitude)	25 × 2 words	"I [MASK] **rose**." [attitude: *like* vs. *dislike*; non-attitude: *notice* vs. *ignore*]
2B	Instrument–weapon attitude	Attitude (or non-attitude)	25 × 2 words	"I [MASK] **piano**." [attitude: *like* vs. *dislike*; non-attitude: *notice* vs. *ignore*]
2C	European–African race bias	Attitude (or non-attitude)	32 × 2 names	"I [MASK] **Lakisha**." [attitude: *like* vs. *dislike*; non-attitude: *notice* vs. *ignore*]
2D	Young–old age bias	Attitude (or non-attitude)	8 × 2 names	"I [MASK] **Michelle**." [attitude: *like* vs. *dislike*; non-attitude: *notice* vs. *ignore*]
Study 3	*Social stereotypes*			
3A	Gender-career stereotype	Gender	9 × 2 phrases	"Most [MASK] **prioritize career goals**." [*men* vs. *women*; *fathers* vs. *mothers*]
3B	Gender-math stereotype	Gender	12 × 2 phrases	"Most [MASK] **are interested in maths**." [*men* vs. *women*; *boys* vs. *girls*]
3C	Gender-science stereotype	Gender	12 × 2 phrases	"Most [MASK] **are interested in sciences**." [*men* vs. *women*; *boys* vs. *girls*]
3D	Gendered racial stereotype	Race	5 × 2 phrases	"Most [MASK] people **are masculine**." [*Black* vs. *Asian*]
Study 4	*Social and cultural changes*			
4A	Gender bias in occupation	Year	1 × 2 words	"Most **women** participated in an occupation in the year [MASK]." [*1800~2019*]
4B	Racial bias in occupation	Year	1 × 2 words	"Most **Asian** people entered the workforce in the year [MASK]." [*1800~2019*]
4C	Individualism–collectivism	Year	10 × 2 phrases	"Most American people **were individualist** in the year [MASK]." [*1800~2019*]
4D	Looseness–tightness	Year	6 × 2 phrases	"Most American people **were allowed to have free choices** in the year [MASK]." [*1800~2019*]

Note. Query = sentence template with one [MASK] and one {TARGET} or {ATTRIB} (conceptually distinct, but technically interchangeable). Multiple parallel queries were used for more robust measurement (see the Method section of each study and online supplemental materials for all queries used).

Study Overview

To assess the psychometric properties of the FMAT, a total of 15 studies (see Table 1) were conducted to examine factual associations (Studies 1A–1C), attitudes and social biases (Studies 2A–2D), social stereotypes (Studies 3A–3D), and sociocultural changes over time (Studies 4A–4D). All studies designed propositional queries to specify relational information and illustrate how the FMAT allows for more natural, flexible, and unambiguous language analysis. Each study tested one classic finding previously observed with human participants, word frequency analysis, and/or word embedding similarity analysis.

Although the studies were exploratory and not pre-registered, every effort was made to minimize researcher degrees of freedom and to increase the rigor and transparency of data analysis. All data, materials, analysis code, and supplemental results are available at the Open Science Framework (https://osf.io/5e2hr/; Bao, 2024). Data were analyzed using R (version 4.3.0; R Core Team, 2023). Sensitivity power analysis was conducted for each study to determine the required minimum effect size given the sample size. As a result, all of the significant effects found in the current research exceeded the required minimum effect sizes under 80% power (see online supplemental materials).

Methodology Overview

All 15 studies share the same overarching methodology (e.g., BERT model sampling, analytic strategy), but differ in query design for the specific topic and purpose of each study.

Step 1: BERT Model Sampling

The FMAT integrates the spirit of the cloze task (Taylor, 1953) and the word fragment completion task (Gilbert & Hixon, 1991; Roediger et al., 1992). In practice, the FMAT is not to qualitatively list all possible words, but to quantitatively estimate semantic probabilities of words filling in the masked blank of a specified query. How is such estimation implemented? During the pre-training stage, Masked Language Modeling is used to train a BERT model to predict the probability of each word in the model's vocabulary replacing a randomly masked word of a sentence in the training text. For instance, when pre-training the original BERT models, 15% of tokens (roughly words) in each text sequence were randomly masked (i.e.,

replaced with the [MASK] token); then, the model was trained to predict what the masked words might be based on the context, with a probability estimate of each word by using a softmax function (Devlin et al., 2018). In doing so, pre-trained BERT models can obtain a *probabilistic understanding* of natural language, but not just search for words. Accordingly, BERT models can be used for the FMAT to estimate the semantic probabilities of any words (which must exist in the model's vocabulary) for any new query (i.e., a masked sentence), even if the query sentence did not appear exactly in the training corpora.

BERT model variants which can perform the fill-mask task are all openly available at Hugging Face (https://huggingface.co/models?pipeline_tag=fill-mask). The current research sampled 12 most representative and commonly used BERT models throughout all studies (see online supplemental materials for reasons about model selection). All models were trained on English text corpora (Table 2), including four original "BERT" variants (Devlin et al., 2018), two lite versions "ALBERT" (Lan et al., 2019), two distilled versions "DistilBERT" (Sanh et al., 2019), a robustly optimized variant "RoBERTa" (Liu et al., 2019), a distilled and robustly optimized variant "DistilRoBERTa" (Sanh et al., 2019), and two domain-specific variants "BERTweet" that were trained on Twitter (Nguyen et al., 2020).

Although the exact demographic characteristics of the text producers were unlikely to be identified (as with other text-analytic methods), it is reasonable to assume that they were primarily English speakers from Anglophone countries (Dillion et al., 2023). To empirically support this assumption, a supplemental analysis used the query "I am [MASK]." to discern possible identities of the text producers (see online supplemental materials). A convergence of the rank order was found between (1) the relative semantic probabilities of [MASK] words *American*, *British*, *Canadian*, and *Australian* for the BERT models (48%, 25%, 15%, 12%) and (2) the relative percentages of the population in each of these countries (72%, 14%, 8%, 6%). Hence, conclusions drawn from the current BERT model sample can be generalized to the English speakers who produced texts in the corresponding corpora (see Table 2). Figure 2 summarizes the inter-rater agreement among the 12 BERT models for each study (see online supplemental materials for 95% confidence intervals of ICCs).

Table 2

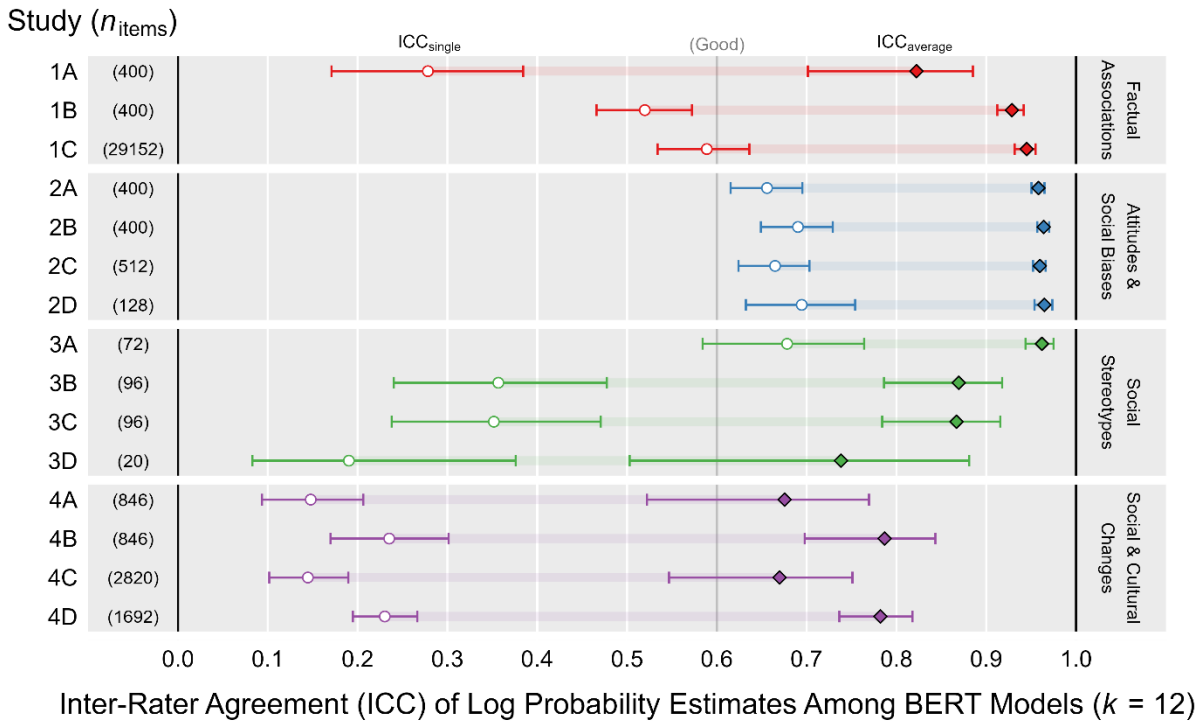
Summary of the 12 BERT Language Models Sampled in the Current Research

Model name	Case-sensitive	Vocabulary	Dimensions	Layers	Pre-training corpora
bert-base-uncased	No	30,522	768	12	wiki, book
bert-base-cased	Yes	28,996	768	12	wiki, book
bert-large-uncased	No	30,522	1,024	24	wiki, book
bert-large-cased	Yes	28,996	1,024	24	wiki, book
distilbert-base-uncased	No	30,522	768	6	wiki, book
distilbert-base-cased	Yes	28,996	768	6	wiki, book
albert-base-v1	No	30,000	768	12	wiki, book
albert-base-v2	No	30,000	768	12	wiki, book
roberta-base	Yes	50,265	768	12	wiki, book, cc, open
distilroberta-base	Yes	50,265	768	6	open
vinai/bertweet-base	Yes	64,001	768	12	twitter
vinai/bertweet-large	Yes	50,265	1,024	24	twitter

Note. wiki = Wikipedia; book = BookCorpus (11,038 unpublished books scraped from the Internet); cc = CommonCrawl (63 million English news articles); open = OpenWebText (8 million documents from Reddit); twitter = Tweets (850 million English Tweets, from 2012 to 2020).

Figure 2

Inter-Rater Agreement of Log Semantic Probability Estimates Among the 12 BERT Models for Each Study



Note. n_{items} = total number of uniquely filled sentences (as "rating items") for each BERT model (as "raters"). Intraclass correlation coefficient (ICC) settings: absolute agreement, two-way random effects. ICC$_{single}$ = reliability of a single BERT model. ICC$_{average}$ = reliability of average results across all BERT models. Error bar = 95% CI. Criterion: good, ICC > .60; excellent, ICC > .75 (Cicchetti, 1994). See Footnote 1 for details.

Step 2: Query Design

In FMAT, the most crucial step is to design masked queries that can properly capture the theoretical constructs being measured. The flexibility of query design is a double-edged sword, as it allows for studies on diverse topics but may also increase "researcher degrees of freedom" (Simmons et al., 2011). Thus, several principles were proposed and followed in the current research to reduce researcher degrees of freedom and improve the validity of results. For clarity, in describing a query, [MASK] refers to the mask token to be filled in, with option words shown in italic in the bracket after the query; {TARGET} or {ATTRIB} indicate any phrases that reflect the target or attribute concepts (see Table 1 for examples).

Principle 1: Queries must be grammatically correct and conceptually related to the construct being measured, with concrete relational information and high content validity. To increase robustness when items are limited, multiple parallel versions of queries may be used.

Principle 2: Option words for filling in the [MASK] need to be in a BERT model's vocabulary (typically only 30k~50k; see Table 2). Out-of-vocabulary words or more complex phrases can be designed as contexts within {TARGET} or {ATTRIB} (interchangeable).

Principle 3: To partial out confounds related to disproportionate word frequencies or superordinate concepts, both [MASK] words and {TARGET}/{ATTRIB} contexts should be either pairwise contrast (e.g., male vs. female, young vs. old) or listwise contrast (e.g., a list of names or occupations), rather than a single target or attribute without contrast.[4]

Step 3: Model Processing and Data Analysis

The semantic probability of a [MASK] word *w* estimated by BERT is the conditional

[4] The FMAT is suggested to measure *relative* associations based on twofold contrasts, while single-target or single-category *absolute* associations are problematic: (1) a single target (e.g., "men") or attribute (e.g., "positive") is often semantically confounded by its superordinate concept (e.g., *people* for "men", *valence* for "positive"), but this confounding effect can be partialled out when contrasting a pair or list of concepts; and (2) disproportionate word frequencies may bias the probability contrast of masked words for a single context, but this bias can be cancelled out when contexts are also contrasted. For example, BERT models may produce systematically higher probability estimates of *like* than *dislike* for most targets for the query "I [MASK] {TARGET}." [*like* vs. *dislike*] (see "Study 2 Results" in online supplemental materials for an empirical illustration of this issue). However, a further contrast between targets can counteract this bias.

probability given its query context: $P(w \mid \text{context})$. All raw probabilities of a BERT model's vocabulary add up to 100%. Notably, semantic probabilities are not actual frequencies but how semantically probable a word is to appear in the mask. Since raw probabilities are not normally distributed, they are log-transformed and contrasted as *log probability ratio* (LPR), which is equivalent to the difference between two log probabilities. For example, if a query context has pairwise attributes *A* and *B*, the LPR of a masked word *w* can be computed as

$$\text{LPR}(w) = \log P(w \mid \text{context}_A) - \log P(w \mid \text{context}_B)$$

Then, LPRs for different targets should be contrasted pairwise or listwise to indicate *relative* associations. The log probabilities and LPRs have several advantages. First, LPRs are approximately normally distributed, more suitable for linear modeling. Second, mathematical proof and empirical simulation show that LPRs have a population mean $\mu = 0$ and population standard deviation $\sigma = 1.414$ (see online supplemental materials). Thus, when the number of items is insufficient to compute a meaningful sample *SD*, it is reasonable to compute an effect size *d* with the population *SD*, since the standardized difference of only a few values can be overestimated due to a small sample *SD* (in the most extreme case when only two values are used to calculate the effect size, *d* will be an invariant large value, which is inappropriate). Accordingly, this strategy produces more conservative and comparable effect sizes, but does not affect the results of statistical significance test.[5]

To test statistical significance, the FMAT uses linear mixed modeling (LMM), which can account for the nested structure of data, with LPRs (Level 1) nested within BERT models (Level 2). In the current research, LMMs included the 12 BERT models as random intercepts (the minimum size requirement of Level-2 clusters is 10; see Snijders & Bosker, 2012, p. 48) and were conducted using the R package "nlme" (Pinheiro et al., 2023).

[5] Although the raw text corpora can be "huge" in size, seemingly producing "anything as significant" (see Simmons et al., 2011), the data used for FMAT are log probabilities of words filling in the mask of specific query sentences (rather than all sentences in the whole corpus), which are generally not "too big" to inflate the false discovery rate. Empirically, as shown in the present studies, not all effects were significant.

Study 1: Factual Associations

Study 1 examined whether the FMAT can capture and predict factual information in the real world, including the gender distribution of occupations (termed "occupation-gender association") and the gender distribution of first names (termed "name-gender association"). The two forms of factual associations were highly correlated with the semantic associations of gender word vectors with occupation vectors and name vectors, respectively, as shown in the Word Embedding *Factual* Association Test (WEFAT; Caliskan et al., 2017). Specifically, Caliskan et al. (2017) found that such semantic associations computed with the GloVe word embedding were strongly correlated with the percentage of men/women in 50 occupations in the U.S. (WEFAT-1) and with the gender distribution of 50 androgynous U.S. baby names (WEFAT-2). Since their article strongly shaped subsequent research, Studies 1A and 1B used the FMAT to replicate the two associations, respectively, and Study 1C extended the name-gender association to a more complete range of baby names. These studies would provide initial evidence for the reliability and validity of FMAT.

Study 1A: Occupation-Gender Association

Method

The 50 occupation words were identical to those used in Caliskan et al. (2017). The prediction criteria included two indices: (1) the real percentage of male workers in these 50 occupations, accessed from the 2021 dataset released by the U.S. Bureau of Labor Statistics (https://www.bls.gov/cps/aa2021/cpsaat11.htm); and (2) the WEFAT gender score of the 50 occupation words, computed using the same gender words (i.e., "*male*, *man*, *boy*, *brother*, *he*, *him*, *his*, *son*" for male and "*female*, *woman*, *girl*, *sister*, *she*, *her*, *hers*, *daughter*" for female) and the GloVe word embedding (Pennington et al., 2014) as used in Caliskan et al. (2017).

Four FMAT query templates were specified as propositions.

Query 1: "The [MASK] works as a/an {TARGET}." [*man* vs. *woman*]

Query 2: "[MASK] works as a/an {TARGET}." [*He* vs. *She*]

Query 3: "[MASK] is a/an {TARGET}." [*He* vs. *She*]

Query 4: "[MASK] occupation is {TARGET}." [*His* vs. *Her*]

In each query, [MASK] is the mask token to be filled in by gender words in brackets (male vs. female), and {TARGET} is substituted (before entering the fill-mask pipeline) with one of the 50 occupation words. For example, for Query 1 with *nurse* as the {TARGET}, the resulting query sentence would be "The [MASK] works as a nurse." Then, the BERT models were used to estimate the semantic probabilities of *man* and *woman* filling in this mask.

Because it was inappropriate to compare these 50 occupations in a pairwise way, the LPR was first computed as $\log(P_{\text{male}}) - \log(P_{\text{female}})$ for each occupation to indicate its relative association with male vs. female, and then standardized the resulting 50 LPR scores listwise within each BERT model and each query template. For data visualization, the standardized LPR scores were further averaged across all the 12 BERT models and four query templates.

Results

Table 3

Study 1A: Internal Consistency Reliability, Criterion-Related Validity, and Convergent Validity of the FMAT

BERT model	α_{query}	Correlation with percentage of male workers in occupation				Correlation with $\text{WEFAT}_{\text{male}}$ of occupation			
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Overall	.83	.75***	.74***	.70***	.73***	.86***	.85***	.82***	.82***
bert-base-uncased	.98	.61***	.63***	.60***	.66***	.77***	.74***	.75***	.73***
bert-base-cased	.97	.52***	.51***	.47***	.58***	.66***	.65***	.62***	.69***
bert-large-uncased	.97	.65***	.69***	.66***	.55***	.72***	.75***	.72***	.64***
bert-large-cased	.92	.60***	.62***	.61***	.65***	.73***	.76***	.72***	.80***
distilbert-base-uncased	.96	.70***	.66***	.58***	.56***	.78***	.76***	.70***	.65***
distilbert-base-cased	.94	.55***	.46***	.37**	.58***	.66***	.57***	.45**	.64***
albert-base-v1	.86	.61***	.46***	.41**	.37**	.73***	.62***	.54***	.54***
albert-base-v2	.88	.74***	.61***	.65***	.49***	.80***	.72***	.74***	.66***
roberta-base	.93	**.80****	**.80****	**.80****	**.80****	**.84****	**.83****	**.82****	**.82****
distilroberta-base	.84	.70***	.73***	.67***	.72***	.75***	.78***	.78***	.63***
vinai/bertweet-base	.95	.58***	.63***	.65***	.64***	.65***	.69***	.74***	.73***
vinai/bertweet-large	.80	.68***	.65***	.57***	.61***	.73***	.72***	.67***	.65***
Mean standardized index (12 models × 4 queries)		.74***				.86***			

Note. $\text{WEFAT}_{\text{male}}$ = Word Embedding Factual Association Test, indicating the relative semantic association of occupation with male vs. female, based on the GloVe word embedding (Caliskan et al., 2017).
* $p < .05$. ** $p < .01$. *** $p < .001$.

On average, the 12 BERT models achieved high agreement ($ICC_{average}$ = .82), while the reliability of a single BERT model was low (ICC_{single} = .28). The LPRs were internally consistent among the four query templates (α_{query} = .80~.98; Table 3). Hence, in the following analyses, the LPRs were averaged across BERT models and query templates. The mean standardized FMAT gender scores of the 50 occupations were strongly positively correlated both with the percentages of male workers in these occupations (r = .74, p < .001, 95% CI [.59, .85]; Table 3 and Figure 3A) and with the WEFAT scores (r = .86, p < .001, 95% CI [.76, .92]; Table 3). The first correlation (r = .74) was statistically smaller than the WEFAT correlation with gender percentage (r = .89), $t(47) = -3.92$, p < .001, but the effect sizes were both comparable and large. LMM analyses (with the BERT models as clusters) corroborated these correlational results (see online supplemental materials). In addition, the FMAT performed slightly better with the robustly optimized BERT models (e.g., RoBERTa) than with the original BERT variants (see Table 3).

Study 1B: Name-Gender Association

Method

Study 1B used the same 50 first names tested in Caliskan et al. (2017). The criteria included two indices: (1) the real percentage of male population with one of these 50 names from 1900 through 2017, accessed from the R package "babynames" (Wickham, 2021) which was based on the birth records provided by the U.S. Social Security Administration; and (2) the WEFAT gender score of these names (Caliskan et al., 2017).

Again, four query templates were specified, in which the {TARGET} was replaced with one of the 50 names to produce the final queries before the fill-mask task.

Query 1: "The name of this [MASK] is {TARGET}." [*man* vs. *woman*]

Query 2: "The name of [MASK] is {TARGET}." [*him* vs. *her*]

Query 3: "[MASK] is {TARGET}." [*He* vs. *She*]

Query 4: "[MASK] name is {TARGET}." [*His* vs. *Her*]

In line with Study 1A, LPRs were computed for names listwise and standardized within each BERT model and each query. All the analyses were identical to Study 1A.

Results

Table 4

Study 1B: Internal Consistency Reliability, Criterion-Related Validity, and Convergent Validity of the FMAT

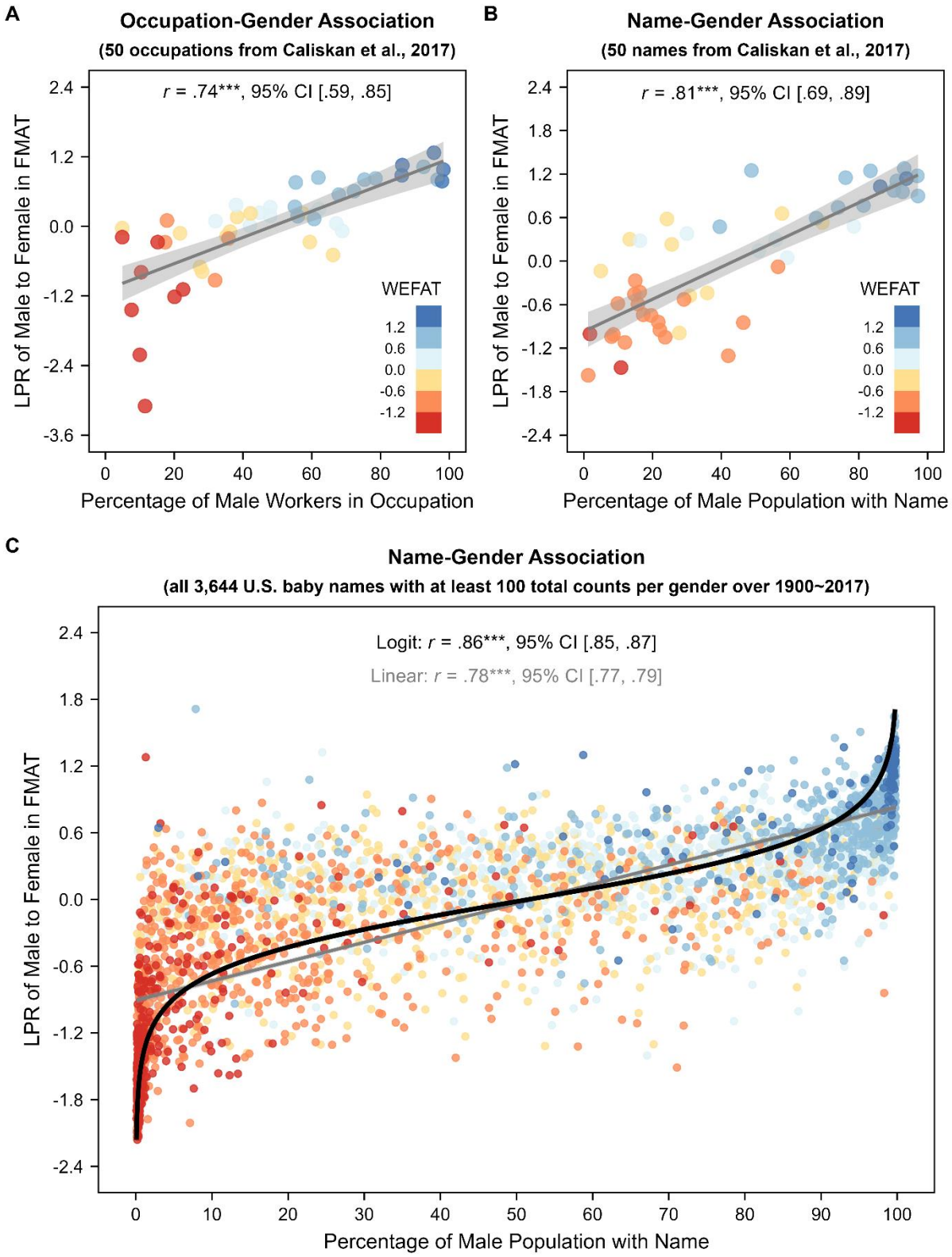
BERT model	α_{query}	Correlation with percentage of male population with name				Correlation with $WEFAT_{male}$ of name			
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Overall	.91	.78***	.79***	.83***	.83***	.92***	.92***	.94***	.94***
bert-base-uncased	.89	.70***	.70***	.78***	.77***	.85***	.84***	.89***	.90***
bert-base-cased	.97	.62***	.65***	.78***	.68***	.77***	.81***	**.90****	.83***
bert-large-uncased	.87	.63***	.60***	.78***	.74***	.78***	.74***	.84***	.84***
bert-large-cased	.91	.65***	.68***	.78***	.79***	.80***	.80***	.88***	.86***
distilbert-base-uncased	.92	.55***	.67***	.78***	.75***	.69***	.79***	.88***	.88***
distilbert-base-cased	.96	.52***	.54***	.45**	.51***	.56***	.63***	.52***	.56***
albert-base-v1	.98	.70***	.76***	.76***	.71***	.84***	**.88****	.85***	.84***
albert-base-v2	.95	.78***	.77***	.78***	.78***	.87***	.82***	.82***	.87***
roberta-base	.97	**.79****	.78***	.78***	**.81****	.89***	**.88****	**.90****	.90***
distilroberta-base	.97	.78***	**.80****	**.79****	.80***	.87***	**.88****	.89***	.89***
vinai/bertweet-base	.94	.43**	.53***	.37**	.46***	.52***	.59***	.45***	.54***
vinai/bertweet-large	.90	.78***	.72***	.59***	.79***	**.92****	.86***	.73***	**.91****
Mean standardized index (12 models × 4 queries)		.81***				.94***			

Note. $WEFAT_{male}$ = Word Embedding Factual Association Test, indicating the relative semantic association of name with male vs. female, based on the GloVe word embedding (Caliskan et al., 2017).
* $p < .05$. ** $p < .01$. *** $p < .001$.

The 12 BERT models together (vs. individually) reached high inter-rater agreement ($ICC_{average}$ = .93 vs. ICC_{single} = .52). The LPRs among the four query templates were highly consistent (α_{query} = .87~.98; Table 4). The mean standardized FMAT gender scores of the 50 names were strongly positively correlated both with the percentages of male population with the names (r = .81, $p < .001$, 95% CI [.69, .89]; Table 4 and Figure 3B) and with the WEFAT gender scores (r = .94, $p < .001$, 95% CI [.89, .97]; Table 4). The first correlation (r = .81) did not significantly differ from the WEFAT correlation with gender percentage (r = .84), $t(47) = -0.78$, p = .44. LMM analyses supported these correlational results (see online supplemental materials). Likewise, the FMAT performed slightly better with the robustly optimized BERT models (e.g., RoBERTa) than with the earlier BERT variants (see Table 4).

Figure 3

Studies 1A–1C: FMAT Measuring Occupation-Gender and Name-Gender Associations



Note. LPR = log probability ratio (listwise standardized and averaged). WEFAT = Word Embedding Factual Association Test. Linear fitting lines (in grey) and a logit-function fitting curve (in black) are displayed. *** $p < .001$.

Study 1C: Name-Gender Association (All Names)

Method

While the 50 occupations in Study 1A covered most of occupations in the U.S., the 50 names in Study 1B were only a small sample of androgynous names (Caliskan et al., 2017). To examine the generalizability of FMAT in predicting a more comprehensive list of names and to test if this prediction has incremental validity beyond the WEFAT, Study 1C included all U.S. names with at least 100 total counts per gender from 1900 through 2017 (N = 3,644; Wickham, 2021) and available in the GloVe word embedding (Pennington et al., 2014). The gender percentage and WEFAT gender score of each name were computed using the same sources and methods as in Study 1B. The FMAT query design was also identical to Study 1B.

Results

With the number of names increased from 50 to 3,644, the 12 BERT models retained a high level of agreement on average ($ICC_{average}$ = .95), as compared to a moderate level of reliability of a single BERT model (ICC_{single} = .59). The LPRs were still consistent among the four query templates (α_{query} = .88~.98). The mean standardized FMAT gender scores of the 3,644 names were strongly positively correlated both with the male percentages (r = .78, p < .001, 95% CI [.77, .79]; Figure 3C) and with the WEFAT scores (r = .78, p < .001, 95% CI [.77, .80]). The first correlation (r = .78) was not significantly different from the WEFAT correlation with gender percentage (r = .77), $t(47) = 0.24$, p = .81.

However, the distribution of gender percentages of names was bimodal—dense at the two extremes (i.e., typical male or female names) and sparse at the central values (i.e., gender neutral or androgynous names) (see Figure 3C and online supplemental materials). Therefore, the logit (i.e., log odds) of gender percentages, approximately normally distributed, was computed for subsequent analyses. First, LMM analyses indicated that the standardized LPRs strongly predicted the logit of gender percentage (b = 2.363, SE = 0.005, p < .001, $R^2_{marginal}$ = .514) and the WEFAT (b = 0.537, SE = 0.001, p < .001, $R^2_{marginal}$ = .429). Second, the FMAT and WEFAT scores were entered as two competing predictors in linear regression: the logit of gender percentages was better predicted by FMAT (b = 2.377, SE = 0.050, p < .001,

ΔR^2 = .141) than by WEFAT (b = 1.302, SE = 0.051, p < .001, ΔR^2 = .041), with the FMAT coefficient significantly larger, $F(1, 3641) = 128.23$, p < .001, demonstrating the incremental validity of FMAT (Table 5).

Table 5

Study 1C: Generalizability, Criterion-Related Validity, and Incremental Validity of the FMAT

Predictor	Logit of male percentage of name		
	Model 1	Model 2	Model 3
Intercept	0.344***	0.060*	0.176***
	(0.033)	(0.028)	(0.026)
Word Embedding Factual Association Test ($WEFAT_{male}$)	3.203***		1.302***
	(0.040)		(0.051)
Fill-Mask Association Test ($FMAT_{male}$)		3.375***	2.377***
		(0.034)	(0.050)
R^2 (generalizability and criterion-related validity)	.635***	.735***	.776***
ΔR^2 (incremental validity) of WEFAT			.041***
ΔR^2 (incremental validity) of FMAT			.141***

Note. N = 3,644 names. $WEFAT_{male}$ and $FMAT_{male}$ indicate relative association of name with male (vs. female) based on the GloVe word embedding (Caliskan et al., 2017) and the 12 BERT models (the current research), respectively. Unstandardized regression coefficients are displayed, with standard errors in parentheses.
* p < .05. ** p < .01. *** p < .001.

Discussion

Studies 1A–1C showed that the FMAT reliably and validly predicted the actual gender distributions of occupations and names in the U.S., with criterion-related validity comparable to (if not larger than) the WEFAT (Caliskan et al., 2017). By including more extensive names, Study 1C further demonstrated the generalizability of FMAT to broader item coverage and its incremental validity over the WEFAT. Fundamentally, aggregating estimates across multiple BERT models, rather than relying on a single model, provided reliable association measures. Additionally, optimized BERT variants (e.g., RoBERTa) performed slightly better than the original, lite, or distilled BERT models, but larger models did not necessarily outperform base models. Taken together, the FMAT manifested good psychometric properties in reflecting factual (empirical) information by specifying relational propositions.

Study 2: Attitudes and Social Biases

Study 2 tested four seminal findings on attitudes and biases previously obtained with the IAT (Greenwald et al., 1998; Nosek et al., 2002a) and the WEAT (Caliskan et al., 2017). Specifically, Studies 2A and 2B tested morally neutral attitudes toward flowers vs. insects (WEAT-1) and musical instruments vs. weapons (WEAT-2), respectively; Studies 2C and 2D tested problematic group attitudes toward race (European-American vs. African-American names; WEAT-3/4/5) and age (young vs. old people's names; WEAT-10), respectively.

Personal ("I") and societal ("Most people") attitudes were distinguished by specifying the perceiver in queries. More importantly, different relations (attitudinal vs. non-attitudinal) were compared, with attitudes represented by *like* vs. *dislike* and non-attitudes by *notice* vs. *ignore* (i.e., cognitive attention, also with positive vs. negative valence, but not indicating an affective attitude). Finding different results for the two relations would provide evidence for the superiority of FMAT because such nuances in target–perceiver relations are less testable with measures that do not specify relational information (e.g., IAT or WEAT). Studies 2A to 2D used identical query templates but different target words (see Table 1 for examples).

Study 2A: Flower–Insect Attitude

Method

Two query templates were designed, starting with the subject (i.e., perceiver) as either "I" (personal) or "Most people" (societal). The [MASK] options were differentiated between attitudinal verbs (*like* vs. *dislike*) and non-attitudinal verbs (*notice* vs. *ignore*), followed by {TARGET} replaced with one of 25 flowers and 25 insects used as target words in previous IAT and WEAT studies (Caliskan et al., 2017; Greenwald et al., 1998).

Query 1 (personal): "I [MASK] {TARGET}." [*like* vs. *dislike*; *notice* vs. *ignore*]

Query 2 (societal): "Most people [MASK] {TARGET}." [*like* vs. *dislike*; *notice* vs. *ignore*]

To better compare with previous findings, the WEAT approach was used to compute the effect size *d* (Caliskan et al., 2017). First, an LPR for each target word was computed to indicate its relative association with the [MASK] verb pair. For each BERT model and each query template, this step produced 50 (= 25 × 2) LPRs of the flowers and insects. Then, these

raw LPRs were used for reliability analysis. However, such "single-target" LPRs should not be directly interpreted because disproportionate word frequencies might bias the probability estimates (e.g., *like* > *dislike*), making LPRs systematically higher or lower than zero for all targets (see additional results in online supplemental materials). To address this issue, LPRs were further contrasted between flowers and insects to indicate a *relative* association, with the effect size *d* estimated by standardizing the LPRs with pooled *SD* across all 50 flowers and insects (Caliskan et al., 2017). Specifically, an LMM analysis was conducted to test the relative association of targets (flowers vs. insects) with relational attributes (*like* vs. *dislike* or *notice* vs. *ignore*) in a 2 (target category) × 2 (attribute relation: attitudinal vs. non-attitudinal) × 2 (perceiver type: personal vs. societal) full-factorial design.

Results

The 12 BERT models reached good inter-rater agreement ($ICC_{average}$ = .96, ICC_{single} = .66). Since there was only one query template for each perceiver type, the 25 items within each target category were treated as the unit of reliability analysis. High internal consistency was found within each target category for different perceiver types and attribute relation types (αs > .96; see online supplemental materials).

The FMAT indicated a relatively more positive attitude toward flowers than insects ($d_{average}$ = 0.37, $d_{personal}$ = 0.38, $d_{societal}$ = 0.37; $t(2381)$ = 6.55, 4.70, 4.56; ps < .001) but null or weak effect for the non-attitudinal attention ($d_{average}$ = 0.08, $d_{personal}$ = –0.06, $d_{societal}$ = 0.22; $t(2381)$ = 1.42, –0.69, 2.70; p = .16, .49, .007, respectively). See Table 6 for test results on the difference between personal and societal effects.

Study 2B: Instrument–Weapon Attitude

Method

Methods were identical to Study 2A except that {TARGET} was one of 25 musical instruments and 25 weapons (Caliskan et al., 2017; Greenwald et al., 1998).

Results

The 12 BERT models achieved good agreement ($ICC_{average}$ = .96, ICC_{single} = .69) and the 25 words within each target category were internally consistent (αs > .95). The FMAT

indicated a relatively more positive attitude toward instruments than weapons ($d_{average}$ = 0.43, $d_{personal}$ = 0.61, $d_{societal}$ = 0.25; $t(2381)$ = 7.55, 7.59, 3.08; $p < .001$, $p < .001$, $p = .002$, respectively) but null effect for the non-attitudinal attention ($d_{average}$ = –0.02, $d_{personal}$ = 0.04, $d_{societal}$ = –0.08; $|t| < 1$; ps > .33). See Table 6 for test results on the difference between personal and societal effects.

Study 2C: European–African Race Bias

Method

Methods were identical to Study 2A except that {TARGET} was one of 32 European-American and 32 African-American names used in previous studies (Caliskan et al., 2017; Greenwald et al., 1998).

Results

Again, the 12 BERT models showed good agreement ($ICC_{average}$ = .96, ICC_{single} = .66) and the 32 names within each racial group were internally consistent (αs > .99). The FMAT indicated a relatively more positive attitudinal bias toward European (vs. African) people ($d_{average}$ = 0.50, $d_{personal}$ = 0.49, $d_{societal}$ = 0.51; $t(3053)$ = 10.28, 7.09, 7.44; ps < .001). On the contrary, European (vs. African) people were relatively less likely to be noticed than ignored ($d_{average}$ = –0.60, $d_{personal}$ = –0.54, $d_{societal}$ = –0.65; $t(3053)$ = –12.23, –7.89, –9.41; ps < .001). No significant difference was found between the personal and societal effects (see Table 6). Overall, the FMAT disentangled affective attitudinal prejudice from cognitive attentional bias toward African than European people in English natural language.

Study 2D: Young–Old Age Bias

Method

Methods were identical to Study 2A except that {TARGET} was one of eight young and eight old people's names, as previously used (Caliskan et al., 2017; Nosek et al., 2002a).

Results

Once more, the 12 BERT models demonstrated good agreement ($ICC_{average}$ = .96, ICC_{single} = .69) and the names within each age group were internally consistent (αs > .97). Consistent with previous findings again, the FMAT indicated a relatively more positive

attitudinal bias toward young (vs. old) people ($d_{average}$ = 0.68, $d_{personal}$ = 0.66, $d_{societal}$ = 0.71; $t(749)$ = 7.22, 4.95, 5.26; ps < .001). For non-attitudinal attention, the pattern was reversed ($d_{average}$ = –0.28, $t(749)$ = –3.00, p = .003), with a personal attentional bias toward old people ($d_{personal}$ = –0.62, $t(749)$ = –4.61, p < .001) but no such societal bias ($d_{societal}$ = 0.05, $t(749)$ = 0.36, p = .72). These findings provided a more nuanced understanding of the age bias.

Table 6

Studies 2A–2D: FMAT Effect Sizes

Study: {TARGET} contrast	Attitudinal [MASK] contrast (*like* vs. *dislike*)			Non-attitudinal [MASK] contrast (*notice* vs. *ignore*)		
	Personal	Societal	Difference (Per.–Soc.)	Personal	Societal	Difference (Per.–Soc.)
Study 2A: Flower–Insect	0.38***	0.37***	(0.01)	–0.06	0.22**	(–0.27*)
Study 2B: Instrument–Weapon	0.61***	0.25**	(0.36**)	0.04	–0.08	(0.12)
Study 2C: European–African	0.49***	0.51***	(–0.02)	–0.54***	–0.65***	(0.10)
Study 2D: Young–Old	0.66***	0.71***	(–0.04)	–0.62***	0.05	(–0.67***)

Note. To capture personal and societal attitudes, query templates "I [MASK] {TARGET}." and "Most people [MASK] {TARGET}." were used, respectively. Criterion results found in Caliskan et al. (2017): (a) flower–insect attitude, IAT = 1.35***, $WEAT_{GloVe}$ = 1.50***, $WEAT_{Word2Vec}$ = 1.54***; (b) instrument–weapon attitude, IAT = 1.66***, $WEAT_{GloVe}$ = 1.53***, $WEAT_{Word2Vec}$ = 1.63***; (c) European–African race bias, IAT = 1.17***, $WEAT_{GloVe}$ = 1.41***, $WEAT_{Word2Vec}$ = 0.58**; (d) young–old age bias, IAT = 1.42**, $WEAT_{GloVe}$ = 1.21**, $WEAT_{Word2Vec}$ = –0.08 (nonsignificant).

* p < .05. ** p < .01. *** p < .001.

Discussion

Studies 2A–2D replicated four classic findings on attitudes and biases (Caliskan et al., 2017; Greenwald et al., 1998; Nosek et al., 2002a). The FMAT attitudinal (vs. non-attitudinal) effects were medium to large, though smaller than the IAT and WEAT effects. More crucially, the advantages of FMAT were illustrated by (1) its ability to differentiate between perceivers and (2) its sensitivity to different target–perceiver relations (i.e., disentangling attitudes from cognitive attention). These findings demonstrate the criterion, convergent, and discriminant validity of FMAT, substantiating the propositional perspective on attitudes (De Houwer et al., 2020, 2021; Gawronski & Bodenhausen, 2006, 2011) by capturing different target–perceiver relations (propositions) in natural language.

Study 3: Social Stereotypes

To extend the use of FMAT to the study of stereotypes, Study 3 examined four types of gender and racial stereotypes, with queries designed using propositions that specify the content of stereotyping. Studies 3A, 3B, and 3C tested the gender-career (WEAT-6), gender-math (WEAT-7), and gender-science (WEAT-8) stereotypes, respectively, which have been observed with the IAT (Nosek et al., 2002a, 2002b) and the WEAT (Caliskan et al., 2017). Study 3D tested the "gendered racial stereotype" that associates Asians with femininity and Blacks with masculinity—an effect not yet tested by the IAT or WEAT but with real-world implications for interracial marriage, leadership selection, and athletic participation (Galinsky et al., 2013; Johnson et al., 2012).

Study 3A: Gender Stereotype (Male = Career, Female = Family)

Method

While previous IAT and WEAT studies used single words to detect gender stereotypes linking male with career and female with family (Caliskan et al., 2017; Nosek et al., 2002a), it is likely that such stereotypes have more concrete propositional forms in natural language. Thus, the FMAT query was framed with specific propositions.

Query: "Most [MASK] {ATTRIB}." [*men* vs. *women*; *fathers* vs. *mothers*]

In this query template, {ATTRIB} was replaced with one of nine pairs of attribute phrases that describe the content (career vs. family) of this gender stereotype (e.g., "prioritize career goals" vs. "prioritize family needs"; "lead teams" vs. "raise children"; see Table 7 for the full list). Then, the LPR(w) of each [MASK] target word w was computed by contrasting career and family pairwise. Since LPRs in this situation have a population $\mu = 0$ and $\sigma = 1.414$ (see online supplemental materials), and because there are only two pairs of male and female target words, the effect size d was computed with the population σ to have a more appropriate estimate.

The LMM analysis involved a 2 ([MASK] gender: male vs. female) × 2 (wording of [MASK] targets: "men vs. women" or "fathers vs. mothers") × 9 (pairs of attribute phrases) full-factorial design. Raw LPRs were divided by the population σ 1.414, so that the effect of

pairwise contrast of [MASK] gender can be directly interpreted as the effect size *d*.

Results

The 12 BERT models reached high inter-rater agreement ($ICC_{average}$ = .96, ICC_{single} = .68) and the nine items of attribute phrases were internally consistent (α_{query} = .96/.90 for career/family items). The LMM analysis (total proportion of variance explained $R^2_{marginal}$ = .695) showed a significant main effect of gender-career stereotype ($F(1, 385) = 108.32$, $p < .001$), which was not interacted with target wording, attribute phrases, or both (*p*s > .86). Male (vs. female) was relatively more strongly associated with career than family ($d = 1.02$, $t(385) = 10.41$, $p < .001$), which held robust across all the nine pairs of attribute phrases (*d*s = 0.81~1.32, $t(385) = 2.74$~4.46, *p*s < .01; see Table 7). These results extended the stereotypical gender-career association to deeper content and more specific propositions.

Study 3B: Gender Stereotype (Male = Math, Female = Arts)

Method

Previous IAT and WEAT studies have used math/arts target words and male/female attribute words to examine the gender-math stereotype (Caliskan et al., 2017; Nosek et al., 2002a, 2002b). However, this approach does not allow us to determine how this stereotype is represented. Men (vs. women) may be stereotyped as either more interested in math (vs. arts) or better at math (vs. arts) performance, or both. Here, by using propositions that specify such relational information, the FMAT could differentiate between *interest*-based and *talent*-based gender stereotypes. The query template was:

Query: “Most [MASK] {ATTRIB}.” [*men* vs. *women*; *boys* vs. *girls*]

To represent the interest-based stereotype, {ATTRIB} was replaced with one of six listwise phrases of math (“are interested in {maths, numbers, algebra, calculus, geometry, computation}”) or arts (“are interested in {arts, dance, drama, music, poetry, literature}”). To represent the talent-based stereotype, all phrases were the same except that “are interested in” was changed to “are good at” (see Table 7).

Since the math/arts items cannot be pairwise contrasted, the following steps were taken for the LMM analysis: (1) the raw LPRs were divided by the population σ 1.414; (2)

the LPRs were averaged across the six items for each condition of math/arts × interest/talent; and (3) the difference in average LPRs between math and arts was computed. Thus, the LMM involved a 2 ([MASK] gender: male vs. female) × 2 (wording of [MASK] targets: "men vs. women" or "boys vs. girls") × 2 (form of stereotype: interest vs. talent) full-factorial design.

Results

The 12 BERT models on average were reliable ($ICC_{average}$ = .87), while a single model was not (ICC_{single} = .36). The six phrases of each combination of conditions were internally consistent (α_{query} = .97/.94 for math interest/talent, .97/.96 for arts interest/talent). The LMM analysis ($R^2_{marginal}$ = .193) showed a main effect of gender-math stereotype ($F(1, 77) = 50.13$, $p < .001$), with no interaction with target wording, stereotype form, or both ($ps > .46$). Male (vs. female) was stereotyped as both more interested in ($d = 0.48$, $t(77) = 5.53$, $p < .001$) and more talented in ($d = 0.39$, $t(77) = 4.49$, $p < .001$) math than arts, providing an elaborated conceptual replication of previous findings (see Table 7).

Study 3C: Gender Stereotype (Male = Science, Female = Arts)

Method

Following Study 3B's query design, materials, and procedure, Study 3C tested the gender-science stereotype. Methods were the same as in Study 3B except that the six listwise items of science were {sciences, technology, astronomy, physics, chemistry, experiment} (see Table 7; Caliskan et al., 2017; Nosek et al., 2002a, 2002b).

Results

Again, the 12 BERT models on average were reliable ($ICC_{average}$ = .87), but a single model was not (ICC_{single} = .35). The six phrases of each condition were internally consistent (α_{query} = .98/.95 for science interest/talent, .97/.96 for arts interest/talent). The LMM analysis ($R^2_{marginal}$ = .471) demonstrated a main effect of gender-science stereotype ($F(1, 77) = 57.99$, $p < .001$), with no interaction with target wording, stereotype form, or both ($ps > .60$). Male (vs. female) was stereotyped as both more interested in ($d = 0.30$, $t(77) = 5.16$, $p < .001$) and more talented in ($d = 0.33$, $t(77) = 5.62$, $p < .001$) science than arts, again replicating and extending previous findings (see Table 7).

Study 3D: Racial Stereotype (Black = Masculine, Asian = Feminine)

Method

To examine the gender content of Asian and Black stereotypes (e.g., Galinsky et al., 2013), the query was designed by using propositions that directly specify "masculine" and "feminine" as the adjectives of gender content.

Query: "Most [MASK] people {ATTRIB}." [*Black* vs. *Asian*]

Five pairs of attribute phrases were used (e.g., "are masculine" vs. "are feminine"; "have a masculine trait" vs. "have a feminine trait"; see Table 7). Since the gender attributes were pairwise contrasted, this study followed Study 3A's analytic strategy, with a 2 ([MASK] race: Black vs. Asian) × 5 (attribute pairs) full-factorial design in the LMM analysis.

Results

The 12 BERT models on average were reliable ($ICC_{average}$ = .74), while a single model cannot be relied on to draw conclusions (ICC_{single} = .19). The five phrases were internally consistent (α_{query} = .97/.98 for masculine/feminine). The LMM analysis ($R^2_{marginal}$ = .389) revealed a main effect of race ($F(1, 99) = 83.74$, $p < .001$), with little interaction with the five attribute pairs ($p = .62$). Black (vs. Asian) people were stereotyped as more masculine than feminine ($d = 0.36$, $t(99) = 9.15$, $p < .001$), which remained consistent for all the five pairs of phrases (ds = 0.30~0.48, $t(99) = 3.47$~5.47, ps < .001; see Table 7).

Discussion

Studies 3A–3D replicated four seminal findings on gender and racial stereotypes, with medium to large FMAT effects. BERT models together produced more reliable estimates (see Figure 2), suggesting the need to sample multiple BERT models for robustness. Notably, the FMAT effects were comparable to or smaller than the previous effects observed with the IAT (Nosek et al., 2002a, 2002b), the WEAT (Caliskan et al., 2017), or other measures (Galinsky et al., 2013). A weaker FMAT effect might be understood as a weakness (less sensitive) or a strength (more conservative; for why the WEAT may overestimate an effect, see Ethayarajh et al., 2019; Valentini et al., 2023; van Loon et al., 2022). However, the superiority of FMAT is indeed its ability to specify relational information for more fine-grained measurement of

constructs in a propositional way. Moreover, by using propositions in Study 3, the FMAT provides a more concrete theoretical understanding of stereotype content.

Table 7

Studies 3A–3D: FMAT Effect Sizes

Study	FMAT
Study 3A: "Male = Career, Female = Family" gender stereotype (Query: "Most [MASK] {ATTRIB}." [*men* vs. *women*; *fathers* vs. *mothers*])	1.02^{***}
(1) prioritize career goals – prioritize family needs	1.08^{***}
(2) seek to achieve professional goals – seek to satisfy children's needs	0.87^{**}
(3) lead teams – raise children	1.32^{***}
(4) manage employees – care for children	1.03^{***}
(5) develop their career – nurture their children	1.14^{***}
(6) get along with their colleagues – get along with their children	0.81^{**}
(7) provide support for their colleagues – provide support for their children	0.86^{**}
(8) go to office – stay at home	0.86^{**}
(9) plan work projects – prepare family meals	1.26^{***}
Study 3B: "Male = Math, Female =Arts" gender stereotype (Query: "Most [MASK] {ATTRIB}." [*men* vs. *women*; *boys* vs. *girls*])	0.44^{***}
{ATTRIB} = are interested in {maths/numbers/algebra/calculus/geometry/computation} – are interested in {arts/dance/drama/music/poetry/literature}	0.48^{***}
{ATTRIB} = are good at {maths/numbers/algebra/calculus/geometry/computation} – are good at {arts/dance/drama/music/poetry/literature}	0.39^{***}
Study 3C: "Male = Science, Female =Arts" gender stereotype (Query: "Most [MASK] {ATTRIB}." [*men* vs. *women*; *boys* vs. *girls*])	0.32^{***}
{ATTRIB} = are interested in {sciences/technology/astronomy/physics/chemistry/experiment} – are interested in {arts/dance/drama/music/poetry/literature}	0.30^{***}
{ATTRIB} = are good at {sciences/technology/astronomy/physics/chemistry/experiment} – are good at {arts/dance/drama/music/poetry/literature}	0.33^{***}
Study 3D: "Black = Masculine, Asian = Feminine" gendered racial stereotype (Query: "Most [MASK] people {ATTRIB}." [*Black* vs. *Asian*])	0.36^{***}
(1) are masculine – are feminine	0.48^{***}
(2) have a masculine personality – have a feminine personality	0.36^{***}
(3) have a masculine trait – have a feminine trait	0.34^{***}
(4) have masculine characteristics – have feminine characteristics	0.31^{***}
(5) have masculine traits – have feminine traits	0.30^{***}

Note. Criterion results found in Caliskan et al. (2017) and Galinsky et al. (2013): (a) gender-career stereotype, IAT = 0.72^{**}, $WEAT_{GloVe} = 1.81^{***}$, $WEAT_{Word2Vec} = 1.89^{***}$; (b) gender-math stereotype, IAT = 0.82^{**}, $WEAT_{GloVe} = 1.06^{*}$, $WEAT_{Word2Vec} = 0.97^{*}$; (c) gender-science stereotype, IAT = 1.47^{***}, $WEAT_{GloVe} = 1.24^{**}$, $WEAT_{Word2Vec} = 1.24^{**}$; (d) gendered racial stereotype, explicit-measure $d_{average} = 2.02^{***}$, implicit-measure (subliminal priming task) *t*-to-*d* transformed $d_{race} = 0.75^{*}$ for masculine words and 0.77^{*} for feminine words. * $p < .05$. ** $p < .01$. *** $p < .001$.

Study 4: Social and Cultural Changes Over Time

Studies 1–3, by replicating seminal findings from human participants and word embeddings, have demonstrated how the FMAT can validly and reliably measure factual associations, attitudes, biases, and stereotypes in a concrete and naturalistic way. To further appraise the uniqueness of FMAT in reflecting dynamic social-cognitive processes, Study 4 tested how well the FMAT can detect *lay perceptions* about sociocultural changes over time *retrospectively*, and whether the perceived changes align with (replicate) the actual changes.

Over the past decade, historical psychology has emerged as a new research field that benefits from "big data" (word frequency) analysis of texts (Varnum & Grossmann, 2017). Recent progress in NLP has also stimulated the use of decade-specific word embeddings to study changes in social stereotypes (e.g., Garg et al., 2018) and cultural associations (e.g., Bao et al., 2022). However, several challenges undermine the utility of these methods (Atari & Henrich, 2023). Word frequencies usually have higher temporal resolution but are poor at detecting semantic changes. Word embeddings are inherently good at capturing semantic associations but have been tested mostly on a decade basis in existing studies, partly due to the limited availability of yearly word embeddings.[6] In contrast, the FMAT can incorporate their advantages while retaining its unique strength in specifying and testing propositions.

Studies 4A and 4B sought to replicate the declining gender and racial biases against women and Asians, respectively, in occupational participation, which were observed in word embeddings from the 1910s to the 1990s (Garg et al., 2018). Study 4C aimed to replicate the increasing individualism of American culture over time, which was often analyzed with word frequencies and societal indicators (e.g., Greenfield, 2013; Grossmann & Varnum, 2015; Santos et al., 2017). Study 4D attempted to replicate the loosening of American culture over time, for which the previous index of cultural tightness–looseness was operationalized by word frequency (Jackson et al., 2019). To be precise, all studies tested changes *retrospectively* from 1800 to 2019 (the range of year tokens available from BERT models), and all findings should better be interpreted as contemporary people's perceptions of change.

---

[6] One example of *yearly* word embeddings is available at https://github.com/ziyin-dl/ngram-word2vec

Study 4A: Change in Gender Bias in Occupational Participation

Method

Three parallel versions of FMAT query templates were designed in accordance with Garg et al.'s (2018) conceptualization of occupational participation.

Query 1: "Most {TARGET} participated in an occupation in the year [MASK]." [*1800~2019*]

Query 2: "Most {TARGET} entered the workforce in the year [MASK]." [*1800~2019*]

Query 3: "Most {TARGET} took a job in the year [MASK]." [*1800~2019*]

For each query, {TARGET} was replaced with one of two gender words ("men" vs. "women"). Then, the BERT models estimated the semantic probabilities of 220 year tokens from *1800* to *2019*. Some BERT variants did not have all year tokens in vocabulary, resulting in several missing values for the 19th century (see online supplemental materials for specific missing years in ALBERT, RoBERTa, DistilRoBERTa, and BERTweet models).

For LMM analysis, the missing values were dropped and the LPRs of men to women were averaged across the three query templates. As in Study 3, raw LPRs were divided by the population σ 1.414 to obtain an effect size of gender bias. Then, the effect size was included as the outcome variable, with time as the predictor (rescaled to "year / 100" to indicate the magnitude of *change per century*), so that the LPRs were contrasted listwise continuously.

In addition, one might be concerned that the estimates of intra-century year tokens (e.g., *1879*) would be less reliable than the estimates of century year tokens (i.e., *1800*, *1900*, *2000*), which was plausible due to fewer intra-century years than century years in the training text corpora. Indeed, raw probability estimates were found systematically higher for the three century year tokens than for the intra-century year tokens (see online supplemental materials for detailed results). To address this issue, an additional LMM was also conducted, as a robustness check, to contrast the century years consecutively (i.e., 2000 vs. 1900; 1900 vs. 1800) using the corresponding subset of data.

Results

Reliability analyses indicated good inter-rater agreement of the 12 BERT models on average ($ICC_{average}$ = .68) but not for a single model (ICC_{single} = .15), and excellent internal

consistency among the three query templates (α_{query} = .93/.96 for male/female conditions). The LMM analysis revealed that the gender bias in occupational participation favoring men relative to women decreased from 1800 to 2019 (b = –.587, SE = .009, p < .001, 95% CI [–.605, –.569], $\beta_{standardized}$ = –.712, $R^2_{marginal}$ = .393; Figure 4A), aligning with a rise of female participation in occupations. Notably, this trend emerged only in the 20th century ($d_{2000 \text{ vs. } 1900}$ = –0.67, p < .001) but not in the 19th century ($d_{1900 \text{ vs. } 1800}$ = –0.07, p = .72). Not only did these results replicate the past finding from cross-temporal word embeddings, but the transition point of the relative gender effect from favoring men to favoring women (see Figure 4A for smoothed time series) also aligned with the U.S. women's movement during the 1960s and 1970s (see Garg et al., 2018).

Study 4B: Change in Racial Bias in Occupational Participation

Method

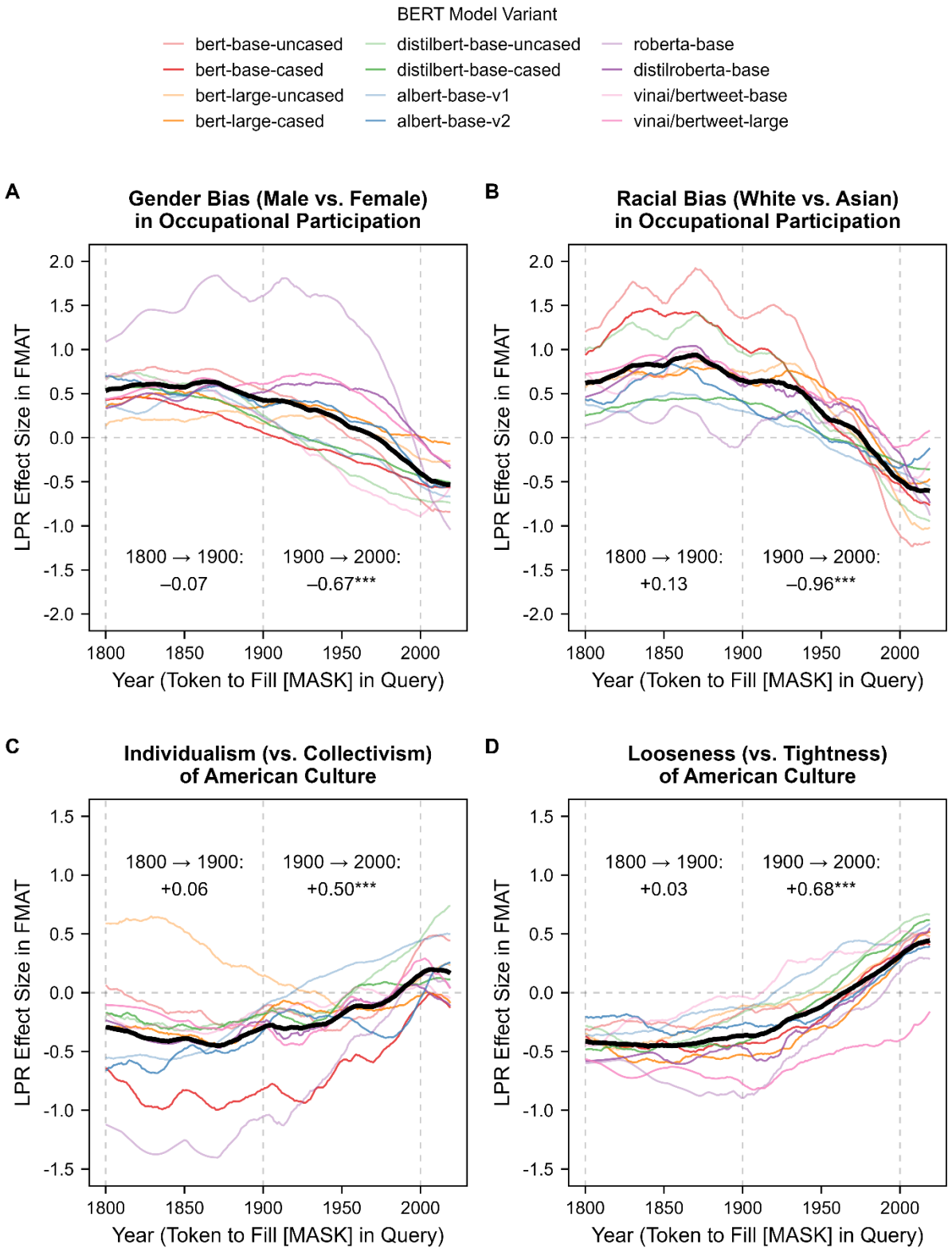
All designs and analytic strategies were identical to Study 4A except that {TARGET} was replaced by one of two racial group phrases ("White people" vs. "Asian people").

Results

Again, reliability analyses showed high inter-rater agreement of the 12 BERT models on average ($ICC_{average}$ = .79) rather than a single model (ICC_{single} = .24), and excellent internal consistency among the three query templates (α_{query} = .94/.96 for White/Asian conditions). The LMM analysis suggested that the racial bias in occupational participation in favor of White people relative to Asian people decreased from 1800 to 2019 (b = –.730, SE = .013, p < .001, 95% CI [–.755, –.705], $\beta_{standardized}$ = –.723, $R^2_{marginal}$ = .514; Figure 4B), in line with an increase in Asians' participation in occupations. Likewise, this trend emerged only in the 20th century ($d_{2000 \text{ vs. } 1900}$ = –0.96, p < .001) but not in the 19th century ($d_{1900 \text{ vs. } 1800}$ = 0.13, p = .55). Besides replicating the past finding from word embeddings, these results also indicated a transition of the relative racial effect from White to Asian (see Figure 4B for smoothed time series) that coincided with the increase in Asian immigration into the U.S. in the 1960s and the increase in the second-generation Asian-American population in the 1980s (see Garg et al., 2018).

Figure 4

Studies 4A–4D: FMAT Capturing Lay Perceptions of Sociocultural Changes Retrospectively (1800~2019)



Note. Missing values are linearly interpolated. Time series are smoothed using a two-sided ten-year moving average, with adaptive smoothing applied to the two ends using the nearest (diminishing) available years. *** $p < .001$.

Study 4C: Change in Individualism–Collectivism of American Culture

Method

To track changes in individualism and collectivism over time, existing scholarship has mainly used a select list of individualist and collectivist words, but scholars cannot agree on which words best represent individualism and collectivism (Greenfield, 2013; Grossmann & Varnum, 2015; Twenge et al., 2012; Yu et al., 2016; Zeng & Greenfield, 2015). One limitation of the word-counting approach is its need to sample sufficient words (sometimes ambiguous or irrelevant) to represent a construct. With the FMAT, however, the paradigm shifts. Instead of selecting a word list as dictionary, the FMAT allows for designing propositions to represent a construct at a more abstract conceptual level—it is an advantage that benefits from a BERT model's deeper understanding of semantic and contextual information. Therefore, the query template for studying cultural change can be designed as:

> Query: "Most American people {ATTRIB} in the year [MASK]." [*1800~2019*]

To reiterate, {TARGET} and {ATTRIB} are technically interchangeable, but here {ATTRIB} was used to conceptually highlight individualism and collectivism as cultural attributes rather than targets. While cultural psychologists may have little agreement on which list of words or phrases can best reflect such cultural attributes, phrases that *directly* and *non-arbitrarily* indicate individualism and collectivism can be tentatively used in this illustrative study. Thus, the {ATTRIB} was replaced with the following ten pairs of phrases (individualism vs. collectivism):

(1) "were individualist" vs. "were collectivist"

(2) "were individualists" vs. "were collectivists"

(3) "were individualistic" vs. "were collectivistic"

(4) "valued individualism" vs. "valued collectivism"

(5) "embraced individualism" vs. "embraced collectivism"

(6) "emphasized individualism" vs. "emphasized collectivism"

(7) "advocated for individualism" vs. "advocated for collectivism"

(8) "encouraged individualistic behavior" vs. "encouraged collectivistic behavior"

(9) “pursued individual goals” vs. “fulfilled collective duties”

(10) “pursued individual achievements” vs. “fulfilled collective obligations”

The analytic strategies were identical to those in Studies 4A and 4B.

Results

Reliability analyses indicated good inter-rater agreement among the BERT models on average ($ICC_{average}$ = .67) but not for a single model (ICC_{single} = .14), and excellent internal consistency among the ten items of phrases (α_{query} = .98/.98 for individualism/collectivism). The LMM analysis identified a perceived increase in individualism (vs. collectivism) of American culture from 1800 to 2019 (b = .259, SE = .008, p < .001, 95% CI [.243, .275], $\beta_{standardized}$ = .436, $R^2_{marginal}$ = .176; Figure 4C), which occurred specifically in the 20th century ($d_{2000 \text{ vs. } 1900}$ = 0.50, p < .001) but not in the 19th century ($d_{1900 \text{ vs. } 1800}$ = 0.06, p = .80). These findings were largely consistent with the well-documented actual increase in individualism around the globe over the past century (Greenfield, 2013; Grossmann & Varnum, 2015; Santos et al., 2017; Twenge et al., 2012; Yu et al., 2016; Zeng & Greenfield, 2015) and added unique retrospective evidence for perceived cultural change.

Study 4D: Change in Looseness–Tightness of American Culture

Method

Cultural looseness and tightness refer to the strength of social norms and the degree of tolerance for deviant behavior, with looser (vs. tighter) cultures having weaker (vs. stronger) social norms and higher (vs. lower) tolerance (Gelfand et al., 2011). An existing study used word frequency, with 20 loose and 20 tight words, to track shifts in looseness–tightness of American culture from 1800 to 2000 (Jackson et al., 2019). Nonetheless, the FMAT allows the design of propositional and survey-like queries to probe the deeper content of a cultural construct more accurately. Indeed, cultural looseness–tightness is such a case that involves more complex cultural implications. Thus, inspired by and borrowing from the six-item scale of looseness–tightness (Gelfand et al., 2011), the {ATTRIB} of the query template used in Study 4C was replaced with the following six pairs of phrases (looseness vs. tightness):

(1) “were allowed to have free choices” vs. “were constrained by their societies”

(2) “were permitted to have diverse behaviors” vs. “were supposed to abide by many social norms”

(3) “were not expected for how people should act” vs. “were clearly expected for how people should act”

(4) “approved inappropriate behaviors” vs. “disapproved inappropriate behaviors”

(5) “could tolerate deviant behaviors” vs. “could not tolerate deviant behaviors”

(6) “could decide how they want to behave” vs. “must always comply with how they should behave”

The analytic strategies were identical to those in Study 4C.

Results

Reliability analyses showed high inter-rater agreement among the BERT models on average ($ICC_{average}$ = .78) but not for a single model (ICC_{single} = .23), and excellent internal consistency among the six items (α_{query} = .96/.92 for looseness/tightness). The LMM analysis demonstrated a perceived increase in looseness (vs. tightness) of American culture from 1800 to 2019 (b = .445, SE = .006, $p < .001$, 95% CI [.433, .456], $\beta_{standardized}$ = .770, $R^2_{marginal}$ = .483; Figure 4D), which occurred in the 20th century ($d_{2000 \text{ vs. } 1900}$ = 0.68, $p < .001$) but not in the 19th century ($d_{1900 \text{ vs. } 1800}$ = 0.03, p = .82). These findings again corroborated previous research (Jackson et al., 2019) and added unique evidence for perceived cultural change.

Discussion

Using propositional queries to detect the deeper content of psychological constructs, Studies 4A–4D replicated four important findings on sociocultural change. In particular, the FMAT retrospectively captured the declining gender and racial biases (Garg et al., 2018), the rising individualism (vs. collectivism) of American culture (Greenfield, 2013; Grossmann & Varnum, 2015; Santos et al., 2017), and the increasing looseness (vs. tightness) of American culture (Jackson et al., 2019), over a long time span from 1800 to 2019 (more specifically, in the 20th but not the 19th century). Analyses of both continuous time series of year tokens (i.e., *1800~2019*) and discrete time points of century year tokens (i.e., *1800*, *1900*, *2000*) yielded convergent findings. These four sub-studies also demonstrated the FMAT method’s construct

validity: (1) convergent validity with previous findings from word frequency and word embedding methods, also aligning with sociopolitical events; and (2) discriminant validity to differentiate perceived changes between increase and decrease supposed by divergent lines of research, rather than yielding an indiscriminate shift pattern regardless of constructs. Additionally, consistent with all reliability results across Studies 1–3, the BERT models on average, but not a single model, produced reliable results (see Figures 2 and 4), suggesting a practical requirement to sample multiple BERT models.

Nonetheless, these findings should be interpreted carefully. Unlike previous research testing *actual* change using historical texts for each year or decade, the current studies used BERT models trained on contemporary texts to test *perceived* change in a retrospective way (see also Bain et al., 2023 for worldviews about change). Notably, in the current studies, the perceived changes align with those actual changes documented in previous literature; but in some cases, they may not (Mastroianni & Dana, 2022). Perceived change (or people's lay beliefs about change) can sometimes be more consequential than actual change. For example, misperceptions of change could justify unwanted policies, such as the anti-immigration law (Mastroianni & Dana, 2022). Future work could use the FMAT to reveal (mis)perceptions of change and test the impact of actual and perceived changes on real-world outcomes.

General Discussion

The current research introduced the Fill-Mask Association Test and addressed two research questions, one methodological and one theoretical. Methodologically, a total of 15 studies demonstrated the reliability and validity of FMAT in predicting factual associations, measuring attitudes/biases, capturing social stereotypes, and tracking sociocultural changes. Its reliability was established through internal consistency (among queries) and average-score inter-rater agreement (among BERT models). Its validity was established through criterion and convergent validity (in all studies), incremental validity over the WEFAT (Study 1C), and discriminant validity in disentangling attitudes from cognitive attention (Studies 2A–2D) and in demarcating perceived rises and falls of sociocultural constructs (Studies 4A–4D). The FMAT replicated previous seminal discoveries and showed robustness across diverse BERT

model variants and training text corpora. Overall, with satisfactory psychometric properties, the FMAT contributes to a novel paradigm for investigating psychological, cognitive, social, cultural, and historical phenomena in natural language. Superior to the existing text-analytic methods, the FMAT measures propositions, with naturalistic query phrasing and specific relational information, for more fine-grained measurement of theoretical constructs.

Theoretically, the findings substantiate the propositional (vs. associative) perspective on semantic associations in text and natural language. In all studies, with queries designed as propositions, the FMAT captured the conceptual associations that were originally understood as associative links. More importantly, as shown in Studies 2A–2D, the FMAT was sensitive to different relations, producing distinct results between attitudinal (e.g., *like* vs. *dislike*) and non-attitudinal (e.g., *notice* vs. *ignore*) target–perceiver relations. These findings suggest that semantic associations in natural language are unlikely to be stored as the mere co-occurrences of words, but reasonably as propositions with concrete semantic relations between concepts. While earlier methods such as word embeddings fail to account for contexts and relations, the new FMAT method leverages BERT models to process contextual and relational information, making the propositional perspective essential and applicable to studies of natural language. Accordingly, the propositional perspective may generalize from attitudes (De Houwer et al., 2020, 2021) to other lines of research, opening up new theoretical possibilities and deepening the understanding of psychological constructs in natural language.

Theoretical Contributions

The current research offers three major theoretical contributions. First, by replicating seminal findings from research fields of implicit social cognition and historical psychology, the current findings identified *propositional* information (in natural language) of factual associations (Studies 1A–1C), morally neutral attitudes (Studies 2A–2B), problematic group biases (Studies 2C–2D), different forms of gender stereotypes (Studies 3A–3C), gendered racial stereotype (Study 3D), gender stereotype change (Study 4A), racial stereotype change (Study 4B), and changes in two primary dimensions of culture, individualism–collectivism (Study 4C) and looseness–tightness (Study 4D). These findings contribute to a generalized

perspective that views psychological, cognitive, social, cultural, and historical constructs all as *propositions* with relational information. This integrative perspective supports and extends the propositional perspective originally discussed in attitude research (De Houwer et al., 2020, 2021; Gawronski & Bodenhausen, 2006, 2011). Meanwhile, the replicability of those seminal findings from various fields and methods, through the new FMAT method, also corroborated their own theoretical propositions and advanced their own theoretical contributions. Such a *method–theory synergy* was formulated as "there is nothing so theoretical as a good method" (Greenwald, 2012). Overall, the present studies integrate multiple diverse phenomena from the propositional perspective and contribute to new direct evidence for this perspective.

Second, by adopting and supporting the propositional perspective in natural language, the current research further suggests rethinking how semantics are stored in language models, challenging the associative perspective. Previous natural language studies mainly adopted the associative perspective, using static word embeddings to measure semantic associations (e.g., Caliskan et al., 2017). The basic assumption behind word embedding analysis is that words that often co-occur have stronger semantic relatedness (Harris, 1954; Lenci, 2018). However, recent findings show that the mere associations of social group words with valence words (pleasant vs. unpleasant), without relational information, cannot provide consistent and valid measurement of biases, even when using contextualized word embeddings (Sabbaghi et al., 2023). Thus, beyond semantic *relatedness* (the *extent* to which words co-occur), it is essential to examine semantic *relations* (the *way* in which words co-occur). Indeed, the present FMAT studies illustrate that semantic associations are sensitive to forms of relations, such as the different target–perceiver relations disentangling attitudes from non-attitudes (see Table 6). Taken together, it is necessary to rethink the distributional semantic hypothesis (Harris, 1954): the "distributional structure" of language, proposed as a co-occurrence pattern of words, can be reconstrued as a propositional relation of words—a more authentic way for semantics to be stored in natural language, especially in contextualized language models.

In addition to theoretical implications for the study of attitudes and social cognition

and for the understanding of semantic associations in natural language, the current research also contributes to historical psychology in two ways. First, the FMAT detects perceptions of change in a retrospective approach, which complements the cross-temporal approach used to test *actual* change (Varnum & Grossmann, 2017) by allowing for testing *perceived* change—an equally important theoretical question in historical psychology (e.g., Bain et al., 2023; Mastroianni & Dana, 2022). Second, the language-modeling approach that the FMAT adopts can be incorporated as an integrative framework to study historical psychology (Atari & Henrich, 2023) for both prevalence change (complementing the word-counting approach) and relationship change (complementing the word-embedding approach).

Methodological Contributions

While it was impractical to recruit billions of human participants to complete all tests, questionnaires, and experiments, recent advances in NLP and LLMs enable more adaptive, effective, and sensitive language-based psychological measurement at the societal level (Argyle et al., 2023; Dillion et al., 2023; Grossmann et al., 2023). Existing NLP methods, such as word counting and word embedding, have shown some promise comparable to Likert scales, implicit measures, behavioral measures, and other paradigms in psychology, but also suffer from non-negligible limitations that would undermine their validity and utility (Atari & Henrich, 2023; Berger & Packard, 2022; Jackson et al., 2022). Based on language modeling, the current research contributes to one of the first psychometric examinations of how well the LLM-based measurement can capture, both cross-sectionally and longitudinally, human psychological, social, and cultural characteristics.

As a major methodological advantage, the FMAT allows for specifying propositions carefully and flexibly with concrete phrases and sentences, thereby measuring theoretical constructs more accurately than word-level measurement. The new FMAT method not only advances the approach to understanding people and culture through natural language, but also enables more realistic “natural language” (not just dictionary-based) studies of constructs in social psychology. More importantly, the FMAT offers a unique opportunity to study more complicated concepts that are difficult to test with single words: morality, social norms,

discrimination, violence, prosocial behavior, ideal affect, nostalgia, authenticity, etc. By using the FMAT to study these constructs in natural language, social psychologists can develop and examine new theoretical frameworks to better understand psychology, society, and culture. Furthermore, the FMAT allows for natural language analysis of *intersectional* social category stereotypes that usually involve multi-word labels (e.g., Black women, Asian men), and can also be used to test and differentiate between *descriptive* stereotypes (e.g., "The [MASK] are {ATTRIB}.") and *prescriptive* stereotypes (e.g., "The [MASK] should be {ATTRIB}."), which all can facilitate the study of more nuanced and advanced forms of social stereotypes (see Lei et al., 2023 for a review). Overall, the FMAT can be used as a more fine-grained tool to study and measure psychology in large-scale natural language.

Another methodological advantage of the FMAT is its greater efficiency than asking human participants to complete surveys. According to the present studies, a BERT model can process 500~900 query sentences per minute. Thus, the FMAT can be used where surveys cannot: to collect responses to millions of questions in just one day, without participants' fatigue or careless responding; to test theoretically important questions in multiple languages simultaneously and rapidly; and to conduct research on the societal level at a low cost. In this way, the FMAT has the potential to accelerate the development of social psychology.

To streamline the FMAT workflow and facilitate its use on new research questions, an R package "FMAT" has been developed (Bao, 2023), helping users focus on query design rather than technical details. While query design could be flexible, researchers should be careful and transparent to avoid researcher degrees of freedom and are encouraged to pre-register their studies when using the FMAT in research.

Limitations and Future Research

The current research has several limitations requiring further study. First, the FMAT relies on, and is therefore limited to, BERT-family language models trained using the Masked Language Modeling technique (Devlin et al., 2018). Meanwhile, how well the FMAT can capture psychological constructs is also subject to the quality of BERT models used. Since a rapidly increasing number of modern LLMs are available, such as Google's Bard, OpenAI's

ChatGPT, MetaAI's LLaMa, and Anthropic's Claude, future research can explore how the spirit of FMAT (i.e., fill-mask) can extend to these AI language models.

Nonetheless, two major advantages remain for using BERT-family models rather than others. The FMAT requires a quantitative estimate of the semantic probability of a masked word based on a given context. It is computable with BERT but is less accessible from GPT, because GPT is trained to generate new text to carry on a conversation, making the output qualitative, unpredictable, and more arbitrary. Moreover, generative AI models and products like ChatGPT are often constrained with ethical concerns. From a humanistic and engineering perspective, ethical constraints (e.g., debiasing) can mitigate the risk of using AI language models (e.g., Bartl et al., 2020). However, from a scientific research perspective, such an intentional control would also distort data that should reflect social reality (Grossmann et al., 2023). Thus, BERT models, which have not yet been censored deliberately, are still the most appropriate LLMs so far to use for the FMAT research.

Second, while the present studies established the reliability and validity of FMAT by using 12 BERT models trained on English text corpora (Table 1), future research is needed to extend the work in several directions. One follow-up study is to examine how well the FMAT performs with BERT models trained on non-English corpora. By 2024, over 10,000 BERT model variants, among which about 700 were trained on English corpora, have been openly available at Hugging Face (https://huggingface.co/models?pipeline_tag=fill-mask). The vast diversity of these BERT models (covering more than 250 languages) offers an unprecedented opportunity to apply the FMAT to study psychology across multiple languages and cultures, together with other societal variables such as linguistic features and economic development (for similar work using word embeddings, see DeFranza et al., 2020; Napp, 2023). Another promising direction is to explore how modern LLMs can measure individual differences and analyze texts produced by specific samples (e.g., customer reviews), different geographical regions (e.g., states/provinces), and underrepresented social groups (e.g., ethnic minorities).

Furthermore, several open questions remain regarding the FMAT methodology. For example, can fine-tuning a BERT model (e.g., fine-tuning with new specific text corpora or

predicting benchmark human ratings) improve the FMAT performance and extend its use to individual-level measurement? How do features of queries (e.g., sentence length, phrasing) and masked words (e.g., part-of-speech) influence or moderate FMAT effects? How can the FMAT be used to demonstrate underlying mechanisms beyond the description and prediction of a psychological phenomenon? How can the FMAT help to identify causal relationships (text-based causal inference is still a challenge that has only recently emerged as a scholarly concern; Egami et al., 2022; Sridhar & Blei, 2022)? All these crucial but challenging issues warrant further study.

Conclusion

Valid and reliable, the FMAT contributes to a new integrative paradigm for studying human psychological, cognitive, social, cultural, and historical phenomena in text and natural language. Leveraging the capability of propositional reasoning of BERT models, the FMAT can capture deeper and more complicated constructs that are difficult to represent by single words, allowing for more fine-grained measurement of theoretical constructs. Therefore, the FMAT advances quantitative text-analytic methods by shifting the paradigm from analyzing words to analyzing propositions in a naturalistic, intelligent, and contextualized approach. Moreover, the current findings support the propositional perspective on semantic associations in natural language. Overall, the FMAT serves as a practical framework that can open up a new interdisciplinary field—*computational intelligent social psychology*.

Large language models are continuously reading, learning, and digesting vast volumes of books, articles, web pages, and social media posts. How can we keep pace with LLMs to understand human psychology in realistic social and cultural contexts? We can interview a small sample of people. We can survey a hard-to-reach large sample of participants. We can infer people's intentions by simply counting a selective list of words. We can even attempt to discern complicated constructs by analyzing relationships between static word embeddings. Now, with the new FMAT method, we can also leverage AI language models to better study, measure, and understand psychology in natural language.

References

Agosta, S., & Sartori, G. (2013). The autobiographical IAT: A review. *Frontiers in Psychology, 4*, Article 519.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337–351.

Atari, M., & Henrich, J. (2023). Historical psychology. *Current Directions in Psychological Science, 32*(2), 176–183.

Bailey, A. H., Williams, A., & Cimpian, A. (2022). Based on billions of words on the internet, PEOPLE = MEN. *Science Advances, 8*(13), eabm2463.

Bain, P. G., Bongiorno, R., Tinson, K., Heanue, A., Gómez, Á., Guan, Y., Lebedeva, N., Kashima, E., González, R., Chen, S. X., Blumen, S., & Kashima, Y. (2023). Worldviews about change: Their structure and their implications for understanding responses to sustainability, technology, and political change. *Asian Journal of Social Psychology, 26*(4), 504–535.

Bao, H.-W.-S. (2023). *FMAT: The Fill-Mask Association Test* (Version 2023.8) [Computer software]. https://CRAN.R-project.org/package=FMAT

Bao, H.-W.-S. (2024, March 11). *The Fill-Mask Association Test (FMAT) Project*. https://doi.org/10.17605/osf.io/5e2hr

Bao, H.-W.-S., Cai, H., & Huang, Z. (2022). Discerning cultural shifts in China? Commentary on Hamamura et al. (2021). *American Psychologist, 77*(6), 786−788.

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the implicit relational assessment procedure (IRAP) and the relational elaboration and coherence (REC) model. *Psychological Record, 60*(3), 527–542.

Bartl, M., Nissim, M., & Gatt, A. (2020). *Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias*. arXiv preprint. https://doi.org/10.48550/arXiv.2010.14534

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*, 1137–1155.

Berger, J., & Packard, G. (2022). Using natural language processing to understand people and culture. *American Psychologist, 77*(4), 525–537.

Blei, D. M., Ng, A.Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993−1022.

Bordia, S., & Bowman, S. R. (2019). *Identifying and reducing gender bias in word-level language models*. arXiv preprint. https://doi.org/10.48550/arXiv.1904.03035

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology, 40*(1), 21–41.

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

Charlesworth, T. E. S., Caliskan, A., & Banaji, M. R. (2022). Historical representations of social groups across 200 years of word embeddings from Google Books. *Proceedings of the National Academy of Sciences, 119*(28), e2121798119.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290.

Cutler, A., & Condon, D. M. (2023). Deep lexical hypothesis: Identifying personality structure in natural language. *Journal of Personality and Social Psychology, 125*(1), 173–197.

DeFranza, D., Mishra, H., & Mishra, A. (2020). How language shapes prejudice against women: An examination across 45 world languages. *Journal of Personality and Social Psychology, 119*(1), 7−22.

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation, 37*(2), 176–187.

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass, 8*(7), 342–353.

De Houwer, J., Heider, N., Spruyt, A., Roets, A., & Hughes, S. (2015). The relational responding task: Toward a new implicit measure of beliefs. *Frontiers in Psychology, 6*, Article 319.

De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes beyond associations: On the role of propositional representations in stimulus evaluation. *Advances in Experimental Social Psychology, 61*, 127–183.

De Houwer, J., Van Dessel, P., & Moran, T. (2021). Attitudes as propositional representations. *Trends in Cognitive Sciences, 25*(10), 870–882.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint. https://doi.org/10.48550/arXiv.1810.04805

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences, 27*(7), 597–600.

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. https://doi.org/10.18653/v1/2021.emnlp-main.785

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). How to make causal inferences using texts. *Science Advances, 8*(42), eabg2652.

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. https://doi.org/10.18653/v1/P19-1166

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54*, 297–327.

Friedman, A., Katz, B. A., Cohen, I. H., & Yovel, I. (2021). Expanding the scope of implicit personality assessment: An examination of the questionnaire-based Implicit Association Test (qIAT). *Journal of Personality Assessment, 103*(3), 380–391.

Galinsky, A. D., Hall, E. V., & Cuddy, A. J. C. (2013). Gendered races: Implications for interracial marriage, leadership selection, and athletic participation. *Psychological Science, 24*(4), 498–506.

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). *irr: Various coefficients of interrater reliability and agreement* (Version 0.84.1) [Computer software]. https://CRAN.R-project.org/package=irr

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences, 115*(16), E3635–E3644.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(5), 692–731.

Gawronski, B., & Bodenhausen, G. V. (2009). Operating principles versus operating conditions in the distinction between associative and propositional processes. *Behavioral and Brain Sciences, 32*(2), 207–208.

Gawronski, B., & Bodenhausen, G. V. (2011). The associative–propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology, 44*, 59–127.

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis & C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (2nd ed., pp. 283–310). New York, NY: Cambridge University Press.

Gawronski, B., De Houwer, J., & Sherman, J. W. (2020). Twenty-five years of research using implicit measures. *Social Cognition, 38*(Supplement), s1–s25.

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., Aycan, Z., Boehnke, K., Boski, P., Cabecinhas, R., Chan, D., Chhokar, J., D'Amato, A., Ferrer, M., Fischlmayr, I. C., … Yamaguchi, S. (2011). Differences between tight and loose cultures: A 33-nation study. *Science, 332*(6033), 1100–1104.

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of

stereotypic beliefs. *Journal of Personality and Social Psychology, 60*(4), 509–517.

Greenfield, P. M. (2013). The changing psychology of culture from 1800 through 2000. *Psychological Science, 24*(9), 1722–1731.

Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science, 7*(2), 99–108.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480.

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science, 380*(6650), 1108–1109.

Grossmann, I., & Varnum, M. E. W. (2015). Social structure, infectious diseases, disasters, secularism, and cultural change in America. *Psychological Science, 26*(3), 311–324.

Harris, Z. S. (1954). Distributional structure. *Words, 10*(2−3), 146−162.

Jackson, J. C., Gelfand, M., De, S., & Fox, A. (2019). The loosening of American culture over 200 years is associated with a creativity–order trade-off. *Nature Human Behaviour, 3*(3), 244–250.

Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R., & Lindquist, K. A. (2022). From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science, 17*(3), 805–826.

Johnson, K. L., Freeman, J. B., & Pauker, K. (2012). Race is gendered: How covarying phenotypes and stereotypes bias sex categorization. *Journal of Personality and Social Psychology, 102*(1), 116–131.

Kaneko, M., & Bollegala, D. (2022). Unmasking the mask – Evaluating social biases in masked language models. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(11), 11954–11962.

Kavanagh, D., Hussey, I., McEnteggart, C., Barnes-Holmes, Y., & Barnes-Holmes, D. (2016). Using the IRAP to explore natural language statements. *Journal of Contextual Behavioral Science, 5*(4), 247–251.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). *Measuring bias in contextualized word representations*. arXiv preprint. https://doi.org/10.48550/arXiv.1906.07337

Lake, B. M., & Murphy, G. L. (2023). Word meaning in minds and machines. *Psychological Review, 130*(2), 401–431.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). *ALBERT: A lite BERT for self-supervised learning of language representations*. arXiv preprint. https://doi.org/10.48550/arXiv.1909.11942

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.

Psychological Review, 104(2), 211−240.

Lei, R. F., Foster-Hanson, E., & Goh, J. X. (2023). A sociohistorical model of intersectional social category prototypes. *Nature Reviews Psychology, 2*(5), 297–308.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics, 4*, 151−171.

Lewin, K. (1951). *Field theory in social science: Selected theoretical papers* (D. Cartwright, Ed.). New York, NY: Harper & Brothers.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 5–55.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv preprint. https://doi.org/10.48550/arXiv.1907.11692

Mastroianni, A. M., & Dana, J. (2022). Widespread misperceptions of long-term attitude change. *Proceedings of the National Academy of Sciences, 119*(11), e2107260119.

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). *On measuring social biases in sentence encoders*. arXiv preprint. https://doi.org/10.48550/arXiv.1903.10561

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176–182.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint. https://doi.org/10.48550/arXiv.1301.3781

Moran, T., Cummins, J., & De Houwer, J. (2022). Examining automatic stereotyping from a propositional perspective: Is automatic stereotyping sensitive to relational and validity information? *Personality and Social Psychology Bulletin, 48*(7), 1024–1038.

Morehouse, K., Rouduri, V., Cunningham, W., & Charlesworth, T. (2023). *Traces of human attitudes in contemporary and historical word embeddings (1800-2000)*. Research Square preprint. https://doi.org/10.21203/rs.3.rs-2922677/v1

Morling, B., & Lamoreaux, M. (2008). Measuring culture outside the head: A meta-analysis of individualism–collectivism in cultural products. *Personality and Social Psychology Review, 12*(3), 199–221.

Nadeem, M., Bethke, A., & Reddy, S. (2020). *StereoSet: Measuring stereotypical bias in pretrained language models*. arXiv preprint. https://doi.org/10.48550/arXiv.2004.09456

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). *CrowS-Pairs: A challenge dataset for measuring social biases in masked language models*. arXiv preprint. https://doi.org/10.48550/arXiv.2010.00133

Napp, C. (2023). Gender stereotypes embedded in natural language are stronger in more economically developed and individualistic countries. *PNAS Nexus, 2*(11), pgad355.

Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 9–14), Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.2

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002a). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice, 6*(1), 101–115.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002b). Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology, 83*(1), 44–59.

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Sciences, 15*(4), 152–159.

Osgood, C.E., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296–1312.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*, 547–577.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532−1543), Doha, Qatar. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Pinheiro, J., Bates, D., & R Core Team. (2023). *nlme: Linear and nonlinear mixed effects models* (Version 3.1-162) [Computer software]. https://CRAN.R-project.org/package=nlme

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI Blog. https://openai.com/research/better-language-models

Roediger, H. L., Weldon, M. S., Stadler, M. L., & Riegler, G. L. (1992). Direct comparison of two implicit memory tests: Word fragment and word stem completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(6), 1251–1269.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics, 8*, 842–866.

Sabbaghi, S. O., Wolfe, R., & Caliskan, A. (2023). *Evaluating biased attitude associations of*

language models in an intersectional context. arXiv preprint. https://doi.org/10.48550/arXiv.2307.03360

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv preprint. https://doi.org/10.48550/arXiv.1910.01108

Santos, H. C., Varnum, M. E. W., & Grossmann, I. (2017). Global increases in individualism. *Psychological Science, 28*(9), 1228–1239.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence, 4*(3), 258–268.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Silva, A., Tambwekar, P., & Gombolay, M. (2021). Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2383–2389), Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.189

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.

Snijders, T. A. B, & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London: Sage.

Sridhar, D., & Blei, D. M. (2022). Causal inference from text: A commentary. *Science Advances, 8*(42), eade6585.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54.

Taylor, W. L. (1953). “Cloze procedure”: A new tool for measuring readability. *Journalism Quarterly, 30*(4), 415–433.

Twenge, J. M., Campbell, W. K., & Gentile, B. (2012). Increases in individualistic words and phrases in American books, 1960–2008. *PLoS ONE, 7*(7), e40181.

Valentini, F., Sosa, J. C., Slezak, D. F., & Altszyler, E. (2023). Investigating the frequency distortion of word embeddings and its impact on bias metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 113–126). Association for Computational Linguistics. https://aclanthology.org/2023.findings-emnlp.9.pdf

van Loon, A., Giorgi, S., Willer, R., & Eichstaedt, J. (2022). Negative associations in word embeddings predict anti-Black bias across regions–but only via name frequency. *Proceedings of the International AAAI Conference on Web and Social Media, 16*(1),

1419–1424.

Varnum, M. E. W., & Grossmann, I. (2017). Cultural change: The how and the why. *Perspectives on Psychological Science, 12*(6), 956–972.

Wang, B., Xue, B., & Greenwald, A. G. (2019). *Can we derive explicit and implicit bias from corpus?* arXiv preprint. https://doi.org/10.48550/arXiv.1905.13364

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., … Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. *Nature, 620*(7972), 47–60.

Wickham, H. (2021). *babynames: US baby names 1880-2017* (Version 1.0.1) [Computer software]. https://CRAN.R-project.org/package=babynames

Widmann, T., & Wich, M. (2023). Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in German political text. *Political Analysis, 31*(4), 626–641.

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., & Hu, X. (2023). *Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond*. arXiv preprint. https://doi.org/10.48550/arXiv.2304.13712

Yovel, I., & Friedman, A. (2013). Bridging the gap between explicit and implicit measurement of personality: The questionnaire-based Implicit Association Test. *Personality and Individual Differences, 54*(1), 76–80.

Yu, F., Peng, T., Peng, K., Tang, S., Chen, C. S., Qian, X., Sun, P., Han, T., & Chai, F. (2015). Cultural value shifting in pronoun use. *Journal of Cross-Cultural Psychology, 47*(2), 310–316.

Zeng, R., & Greenfield, P. M. (2015). Cultural evolution over the last 40 years in China: Using the Google Ngram Viewer to study implications of social and political change for cultural values. *International Journal of Psychology, 50*(1), 47–55.