Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Expanding the methodological toolbox: Machine-based item desirability ratings as an alternative to human-based ratings☆

Björn E. Hommel *

Wilhelm Wundt Institute of Psychology, Leipzig University, Germany
magnolia psychometrics GmbH, Germany

ARTICLE INFO

Keywords:
Social desirability bias
Item desirability
Natural language processing
Large language models
Sentiment analysis
Artificial intelligence

ABSTRACT

The accuracy of self-reported data in the social and behavioral sciences may be compromised by response biases such as socially desirable responding. Researchers and scale developers therefore obtain item desirability ratings, in order to maintain item neutrality, and parity with alternative options when creating forced-choice items. Gathering item desirability ratings from human judges can be time-consuming and costly, with no consistent guidelines with regard to required sample size and composition. However, recent advancements in natural language processing have yielded large language models (LLMs) with exceptional abilities to identify abstract semantic attributes in text. The presented research highlights the potential application of LLMs to estimate the desirability of items, as evidenced by the re-analysis of data from 14 distinct studies. Findings indicate a significant and strong correlation between human- and machine-rated item desirability of .80, across 521 items. Results furthermore showed that the proposed fine-tuning approach of LLMs results in predictions that explained 19 % more variance beyond that of sentiment analysis. These results demonstrate the feasibility of relying on machine-based item desirability ratings as a viable alternative to human-based ratings and contribute to the field of personality psychology by expanding the methodological toolbox available to researchers, scale developers, and practitioners.

1. Introduction

Social desirability bias is a pervasive phenomenon that affects the accuracy of self-reported data in the social and behavioral sciences (e.g., Krumpal, 2013; Nederhof, 1985). Survey respondents are inclined to conceal socially undesirable traits and endorse statements that cast them in a favorable manner. Past research has commonly distinguished between two major facets of social desirability bias: self-deception, which constitutes positively biased responses that subjects believe to be true, and impression management, which refers to deliberate attempts to convey a favorable image to specific audiences (Paulhus, 1986).

Some of the methods proposed to cope with the potential threats of impression management involve creating forced-choice questionnaires with items possessing an equal degree of desirability (e.g., Converse et al., 2010; Hughes, Dunlop, Holtrop, & Wee, 2021; Pavlov, Shi, Maydeu-Olivares, & Fairchild, 2021; Wetzel, Frick, & Brown, 2021). In a similar vein, others have suggested devising instruments purely consisting of items of neutral desirability (e.g., Wood, Anglim, & Horwood, 2022). To this end, a well-established approach for evaluating the desirability of items is employing survey respondents or a panel of judges to rate individual items on a desirability scale (Edwards, 1957, p. 5). However, there are inherent challenges associated with obtaining item desirability ratings from judges. Pavlov et al. (2021) have underscored several important considerations, including determining sample size and its composition (e.g., subject matter experts versus target audiences), as well as the level of generalizability of ratings (i.e., whether they reflect general or context-specific desirability). The authors also note the absence of consistent and definitive guidelines in the existing literature regarding these decisions. Furthermore, from the perspective of scale developers, obtaining item desirability ratings may introduce an additional expensive and time-consuming step to an already lengthy scale development process. For example, in a recent study by Ryan et al. (2021), 157 judges were recruited, trained, and instructed to rate 1470 personality statements for item desirability.

Building upon the challenges of obtaining item desirability ratings from human judges, recent advances in natural language processing and

☆ I have no known conflicts of interest to disclose.

* Department of Work and Organizational Psychology, Wilhelm Wundt Institute of Psychology, Leipzig University, Neumarkt 9-19, 04109 Leipzig, Germany.
E-mail address: bjoern.hommel@uni-leipzig.de.

https://doi.org/10.1016/j.paid.2023.112307
Received 14 April 2023; Received in revised form 7 June 2023; Accepted 11 June 2023
Available online 21 June 2023

deep learning introduce a promising alternative. Large language models (LLMs) have emerged as powerful tools, exhibiting remarkable competence in a range of linguistic tasks. This article demonstrates how LLMs can be modified to judge item desirability with high precision as evidenced by a comparison to data from human raters. This work contributes to the field of personality psychology by expanding the methodological tools available to researchers, scale developers, and practitioners by introducing a computerized alternative to human-based item desirability ratings. A web application demonstrating machine-based item desirability rating is provided on: https://huggingface.co/spaces/magnolia-psychometrics/item-desirability-demo

1.1. Utilizing LLMs to evaluate item desirability

With the introduction of the transformer-model architecture, natural language processing has advanced significantly (for in-depth explanations of deep neural networks and transformer-based LLMs, see Hommel, Wollang, Kotova, Zacher, & Schmukle, 2022, and Urban & Gates, 2021). Transformer-based LLMs have recently demonstrated their utility in psychological research, as scholars have successfully employed LLMs to automatically generate personality items (Götz, Maertens, Loomba, & van der Linden, 2023; Hommel et al., 2022; Lee, Fyffe, Son, Jia, & Yao, 2022), conduct content analysis (Fyffe, Lee, & Kaplan, 2023), and extract psychological information from written text (Fan et al., 2023; van Genugten & Schacter, 2022), among other applications. The success of these models can largely be attributed to their capacity for transfer-learning. Through a pre-training process, LLMs acquire general language knowledge and subsequently gain domain-specific expertise when fine-tuned for more narrowly defined tasks on specific training data, such as judging item desirability.

Sentiment analysis is one domain in which LLMs have demonstrated comparable levels of proficiency to humans. This task usually involves categorizing text into pre-defined labels, based on its valence (i.e., positive, neutral, or negative). For example, sentiment analysis may classify the statement "*I make friends easily*" used in the International Personality Item Pool (Goldberg et al., 2006) to assess individual differences in extraversion as *positive*, with a probability of 79 %. Previous research has established a close association between ratings of valence and item desirability (Britz, Gauggel, & Mainz, 2019; Britz, Rader, Gauggel, & Mainz, 2022). Taken together, it is plausible to expect that with sufficient training data, LLMs can learn to predict item desirability.

It is important to note that the method presented in this article implies that items possess a true score in terms of their perceived desirability. The assumption that item desirability is most adequately represented by averaging individual ratings of judges has recently been challenged by Pavlov et al. (2021), who showed that more balanced forced-choice item blocks can be constructed if disagreements between judges are incorporated in the item-matching procedure. Although the proposed LLM-based method aims to predict item desirability as a point estimate, it should not be misconstrued as conducting a desirability rating study with just a single individual judge, as LLMs encode terabytes of human-generated textual data, including expressions of attitudes and social interactions.

In summary, the potential benefits of employing LLMs for evaluating item desirability are threefold. First, LLMs offer a cost-effective alternative to human-based ratings and the potential of evaluating item desirability on a larger scale. Once fine-tuned for this purpose, machine-based evaluation can be performed inexpensively and quickly, without the need for specialized hardware, yielding results within seconds. Second, an LLM-based point estimate of item desirability implicitly reflects diverse perspectives of human judgments. Finally, LLMs can provide a standardized and consistent approach to evaluating item desirability.

2. Method

Materials, data, and code for the present study are available through the Open Science Framework: https://osf.io/67mkz/. Data pre-processing, model training, and statistical analyses were conducted using Python (version 3.8.13) and R (version 4.2.1).

2.1. Data collection

To explore the predictive capacity of LLMs in determining human-rated item desirability, the study drew on a foundation of previously published data for analysis. Using Google Scholar, PsychINFO, and Web of Science, I conducted a literature search for studies reporting item desirability ratings using each of the keywords listed in the OSF repository accompanying this report. This resulted in a list of 234 peer-reviewed publications, of which 14 provided adequate data (i.e., stimulus material in the form of single adjectives or item stems in English or German, as well as reported mean-rated item desirability) either in manuscript tables or in freely accessible online repositories. An overview of the data included in the present study can be found in Table 1.

2.2. Data pre-processing

To ensure consistency in analyzing the data collected from various studies that employed different rating scales to measure item desirability, I z-transformed the human-rated point estimates, taking into account the specific study and questionnaire from which the data originated. When LLMs evaluate individual units of text (e.g., words), they consider the context in which such units occur (Vaswani et al., 2017). I thus used string interpolation to embed adjectives in the dataset in sentences (e.g., "A person is *gullible*."). Finally, text data was cleaned using the Python clean-text package (Filter, 2018) and spell-checked.

2.3. Models used in this study

All analyses of stimulus material (i.e., adjectives and item text) were based on two modified versions of the twitter-XLM-roBERTa-base model (referred to as the "base model"), an LLM trained by Barbieri, Anke, and Camacho-Collados (2022; based on the roBERTa architecture, as proposed by Liu et al., 2019). Barbieri and colleagues fine-tuned this model for sentiment analysis on a multi-lingual dataset of approximately 198 million tweets, categorized into negative, neutral, and positive sentiment. It is freely accessible from https://github.com/cardiffnlp/xlm-t under the Apache 2.0 license. For any given text input, the model produces a vector with three values indicating the class-membership probabilities for each of the sentiment labels. The two modified versions used in this study are described below. Models were trained using Python using the *transformers* package (Wolf et al., 2020) on a Nvidia GeForce RTX 2070 Super GPU, using the CUDA 9.1.85 and cuDNN 7.6.3 toolkits.

2.3.1. Model for sentiment analysis

As item desirability constituted a continuous variable in the data included in this study, I modified and re-trained the base model for regression, as opposed to classification, according to Fig. 1. In simplified terms, the anatomy of LLMs can be divided into an input layer, a multi-layered body, and a classification head. The body of the base model comprises a 12-layer neural network that preserves the LLM's bulk of knowledge in the form of learned parameters (i.e., model weights and biases). The model head, in turn, is trained on a specific task (i.e., classification of sentiment) where it is fine-tuned to make predictions based on the encoded representations provided by the body. As the base model head was designed to predict class probabilities of three labels, I discarded and replaced it with a layer culminating towards a single neuron to project one continuous variable. Re-connecting the model body with the new regression head required fine-tuning the model on

Table 1
Included studies and data characteristics.

Study	Instrument	Language	*k*	*n*	*M*	*SD*
Anderson (1968)	Anderson's List of Personality-Trait Words	English	555	100	2.93	1.46
Schönbach (1972)	Schönbach's List of Personality-Trait Words	German	100	170	2.73	1.60
Bochner and Van Zyl (1985)	Bochner & Van Zyl's Compilation of Personality-Trait Words	English	110	171	4.04	1.59
Hampson, Goldberg, and John (1987)	Goldberg's Personality-Descriptive Terms	English	572	55	4.80	1.93
Dumas, Johnson, and Lynch (2002)	Dumas' Compilation of Personality-Descriptive Words	English	77	581	3.63	1.68
Chandler (2018)	Anderson's List of Personality-Trait Words	English	1106	39	2.95	1.57
Chandler (2018)	Chandler's Compilation of Personality-Trait Words	English	976	47	2.44	1.26
Andersen and Mayerl (2019)	List of Teacher-Related Characteristics	German	30	77	0.75	1.95
Britz et al. (2019)	Aachen List of Trait Words - German Version	German	1212	100	−0.04	1.68
Hughes et al. (2021)	Big Five Aspects Scale	English	98	42	4.07	1.65
Hughes et al. (2021)	Big Five Inventory 2	English	60	42	4.19	1.78
Hughes et al. (2021)	Five-Factor Markers	English	38	43	4.51	1.72
Hughes et al. (2021)	International Personality Item Pool - NEO	English	239	42	4.01	1.61
Leising, Vogel, Waller, and Zimmermann (2021)	Balanced Inventory of Desirable Responding - German Version	German	20	30	−0.04	0.33
Leising et al. (2021)	Beck Depression Inventory - Modified German Version	German	20	30	−0.52	0.21
Leising et al. (2021)	Big Five Inventory - 44 Items - German Version	German	44	44	0.23	0.48
Leising et al. (2021)	Borkenau & Ostendorf's German Adjectives	German	60	24	0.05	0.58
Leising et al. (2021)	International Personality Item Pool - 120 Items - German	German	120	25	0.01	0.45
Leising et al. (2021)	Interpersonal Adjective List	German	16	30	−0.04	0.63
Leising et al. (2021)	Level of Personality Functioning Scale	German	60	24	−0.09	0.55
Leising et al. (2021)	Level of Personality Functioning Scale - Self-Report	German	80	30	−0.17	0.37
Leising et al. (2021)	Life-Orientation-Test - German Version	German	10	30	0.23	0.48
Leising et al. (2021)	Narcissistic Personality Inventory - German Version	German	80	30	0.11	0.32
Leising et al. (2021)	Rosenberg's Self-Esteem Scale - Revised German Version	German	10	30	−0.01	0.61
Leising et al. (2021)	Social Desirability Scale - 17 Items - German Version	German	17	30	0.18	0.61
Wessels, Zimmermann, and Leising (2021)	Wessels et al.'s Compilation of Life Experiences	German	47	18	5.69	2.26
Britz et al. (2022)	Aachen List of Trait Words - English Version	English	1000	203	0.20	1.61
McIntyre (2022)	Big Five Inventory - 44 Items	English	43	193	4.65	1.64
McIntyre (2022)	O*NET Interest Profiler Short Form	English	60	191	4.68	0.62
McIntyre (2022)	Person-Thing Orientation Scale	English	13	193	4.90	0.66
Wood et al. (2022)	International Personality Item Pool - 50 Items	English	24	73	4.35	2.10
Wood et al. (2022)	International Personality Item Pool - 50 Neutralized Items	English	24	73	4.24	1.57

Note. k = Group-wise item/adjective count; k = Group-wise sample size of judges; *M, SD* = Mean and standard deviation of item desirability ratings.

metric training data using mean squared error (MSE) optimization. To achieve this, I re-scored the original training text data used by Barbieri et al. (2022) and subtracted the class probabilities for negative sentiment from the predictions for positive sentiment (see Fig. 1a to Fig. 1b). As re-training merely served to project the information contained in the model body to the head, I prevented the body's parameters from updating during the training phase by a practice commonly referred to as "freezing layers" (Lee, Tang, & Lin, 2019). Apart from these changes, I followed the procedure described by Barbieri et al. (2022). This modified model (referred to as the "sentiment model") exhibited a near-perfect correlation of .99 with the base model's predictions of the test data supplied by Barbieri et al. (2022).

2.3.2. Model for item desirability analysis

The second model used in this study was based on the sentiment model but further fine-tuned to predict item desirability ratings (referred to as the "desirability model"; see Fig. 1c), using the data sources mentioned above. Employing a k-fold cross-validation approach (k = 10), items were grouped by study and questionnaire, and then randomly assigned to a training, validation, or test set, with an 80–10-10-split probability for each group. Urban and Gates (2021) provide an accessible introduction to k-fold cross-validation. Items and adjectives co-occurring across multiple subsets were only assigned once to a single partition to prevent biasing by the same stimulus being present in multiple partitions. The training partition thus comprised 2740 items and interpolated adjectives with respective item desirability ratings. Fine-tuning terminated after 570 straining steps due to early stopping with an $MSE = 0.36$ for the best-performing fold ($M = 0.41$, $SD = 0.05$).

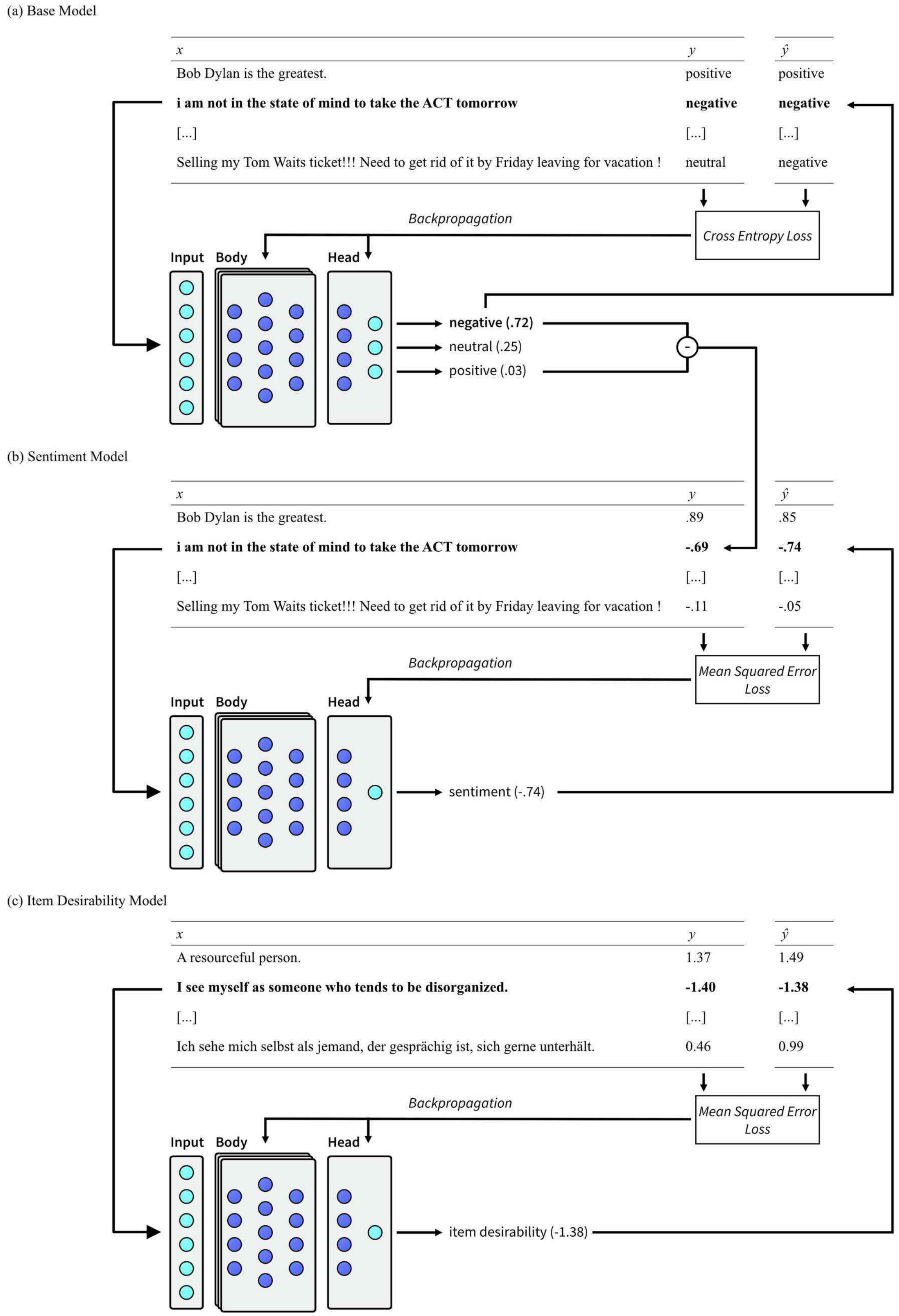
2.4. Measures and covariates

Group-wise z-transformed human-rated item desirability constituted the dependent variable in this study. To predict item desirability as judged by human raters, two machine-based measures were employed; one derived from the sentiment model, and the other from the desirability model. I included three binary covariates in the analysis to assess the accuracy of machine-rated item desirability under more specific circumstances. Personality items, such as the statement "I am very content with myself" (Wood et al., 2022) may be less context-dependent compared to items in other questionnaires, such as occupational interests (e.g., "[…] to create special effects for movies."; Rounds, Su, Lewis, & Rivkin, 2010, as cited in McIntyre, 2022). I thus hypothesized that the former is more easily evaluated by LLMs, yielding a higher convergence between human- and machine-based ratings. I further expected the language of the stimulus material (English versus German) to moderate the prediction, considering the well-documented observation that even multi-lingual LLMs tend to perform better overall for tasks involving English text (Reimers & Gurevych, 2020). Lastly, as LLMs acquire the majority of their knowledge through pre-training on textual data authored by non-psychologists, I anticipated that the predictions of LLMs would align more closely with item desirability judgments made by laypeople rather than those made by psychology students.

3. Results

Analysis conducted on the 521 items in the test and validation set revealed a high level of agreement of ρ = .80 between human- and machine-rated item desirability. These predictions were significantly stronger than compared to machine-rated sentiment ($\rho = .66$, $p < .001$), as determined by Steiger's (1980) test for dependent correlations. Extreme discrepancies between human- and machine-rated item desirability (measured in standardized residuals; $SD \geq |2|$) were observed in 31 items (6 %; see Fig. S1 in the online supplemental material for further details).

I subsequently conducted multiple regression analysis to examine the extent to which the predictive power of the item desirability model



(a) Base Model

x	y	$\hat{y}$
Bob Dylan is the greatest.	positive	positive
i am not in the state of mind to take the ACT tomorrow	**negative**	**negative**
[...]	[...]	[...]
Selling my Tom Waits ticket!!! Need to get rid of it by Friday leaving for vacation !	neutral	negative

(b) Sentiment Model

x	y	$\hat{y}$
Bob Dylan is the greatest.	.89	.85
i am not in the state of mind to take the ACT tomorrow	**-.69**	**-.74**
[...]	[...]	[...]
Selling my Tom Waits ticket!!! Need to get rid of it by Friday leaving for vacation !	-.11	-.05

(c) Item Desirability Model

x	y	$\hat{y}$
A resourceful person.	1.37	1.49
I see myself as someone who tends to be disorganized.	**-1.40**	**-1.38**
[...]	[...]	[...]
Ich sehe mich selbst als jemand, der gesprächig ist, sich gerne unterhält.	0.46	0.99

Fig. 1. Simplified schematic diagram of models and training data used in this study.
Note. Illustration of the basic architecture and training data for (a) the base model for sentiment classification by Barbieri et al. (2022), (b) its modification for regressive sentiment prediction (sentiment model), and (c) the further fine-tuned model for item desirability prediction. Backpropagation updates model parameters for model head and (a, c) body during fine-tuning. y = observed values represented by (a) sentiment classes in original training data, (b) differences between positive and negative class membership probabilities, and (c) human-rated item desirability values; $\hat{y}$ = predicted values by the respective LLM.

Table 2
Results of Linear Regression Analyses for the prediction of human-rated item desirability.

	β	*SE*	*t*	*p*	R^2
Sentiment main effect model					.44
Intercept	0.00	0.03	4.63	<.001	
Machine-rated item sentiment	0.66	0.03	20.2	<.001	
Desirability main effect model					.63
Intercept	0.00	0.03	−3.28	<.001	
Machine-rated item desirability	0.79	0.03	29.63	<.001	
Desirability interaction model					.64
Intercept	0.00	0.05	−4.10	<.001	
Machine-rated item desirability	0.86	0.06	16.86	<.001	
Stimulus content domain	−0.02	0.09	−0.60	.548	
Stimulus language	0.05	0.06	1.57	.116	
Rater group	0.07	0.05	2.66	.008	
Machine-rated item desirability × Stimulus content domain	0.03	0.12	0.90	.369	
Machine-rated item desirability × Stimulus language	−0.06	0.07	−1.79	.073	
Machine-rated item desirability × Rater group	−0.04	0.06	−0.94	.348	

Note. Stimulus content domain (0 = personality, 1 = other), Stimulus language (0 = English, 1 = German), Rater group (0 = laypeople, 1 = psychology students).

varied depending on different covariates. Specifically, I examined possible moderating effects of the content domain (personality versus other) and language (English versus German) of the stimulus material, as well as the rater group (laypeople versus psychology students) who judged item desirability. As shown in Table 2, none of these interactions demonstrated a significant effect, suggesting that the machine-rated item desirability was able to deliver similarly accurate predictions across all conditions examined. Additional variance explained by the moderated model was trivial ($\Delta R^2 = .01$).

4. Discussion

The key finding of this study is a strong Spearman correlation coefficient of .80 between the machine- and human-rated desirability scores, suggesting that the machine model is capable of ranking the estimated desirability of items in a manner that is largely consistent with human judgments. This level of concurrence between the model's predictions and human ratings likely exceeds the consensus among judges in most desirability studies. Results furthermore indicated that the proposed fine-tuning approach of the LLM results in predictions that explained variance beyond that of sentiment analysis. Moreover, the machine prediction of item desirability appears robust for items in the domain of personality, as well as other domains (e.g., occupational interests), and across different languages (i.e., English and German). These predictions do not appear to align more closely with the judgments of laypeople than with those of experts (i.e., psychology students).

This article contributes to the field of personality psychology by broadening the methodological options accessible to researchers, scale developers, and practitioners. In the past, the measurement of item proneness to impression management was confined to the evaluation of stimulus material by human judges. The approach introduced in this article is fundamentally different, as it uses advanced natural language processing techniques to automatically obtain estimates of item desirability in an instant.

The central limitation of this study is that it currently cannot determine the exact circumstances under which a machine model can be used to substitute human judges, as no clear pattern emerges as to how residuals result. In a few cases (6 % of the examined items) extreme discrepancies between human and machine ratings can be observed (e.g., "self-centered"; $\varepsilon = -2.66$; see Fig. S1 in the online supplemental material). A qualitative examination suggests that these exceptional cases arise from a combination of both underfitting (i.e., the estimates reflecting sentiment rather than desirability) and overfitting (i.e. the model becomes excessively specific to the training data). Given the study's restricted quantity and variety of training data (i.e., 2740 items and adjectives originating from low-stakes contexts), this issue can likely be addressed by increasing the amount and diversity of the items in future fine-tuning studies. Additional methodological solutions such as utilizing loss-functions that penalize extreme outliers (e.g., Huber loss; Huber, 1964) and employing regularization (e.g., Urban & Gates, 2021) may be investigated.

Furthermore, as briefly mentioned in the introduction of this article, the assumption that items possess a true desirability score has recently been called into question (Pavlov et al., 2021). The LLM employed in this study predicts item desirability as a point estimate and does not account for the potential heterogeneity of opinion among subsets of judges. The importance of incorporating heterogeneity in perceived desirability is exemplified by the fact that certain personality traits are considered more or less socially desirable across different cultures (Ryan et al., 2021). To address this limitation, future research can explore two avenues. First, apart from point estimates, LLMs could be trained using measures of statistical dispersion. Second, researchers could investigate whether uncertainty measures of the LLM's predictions align with systematic errors in human judgments (e.g., by using Monte Carlo dropout; Gal & Ghahramani, 2016).

Further research may also be dedicated to investigating whether LLM-based estimates yield more generalizable predictions of item desirability compared to desirability ratings obtained from studies with human judges. Such a hypothesis may be justified by the fact that the base model employed in this study was originally trained on an extensive dataset of 2.5 terabytes, comprising filtered text in 100 languages (Liu et al., 2019). It is thus plausible to propose that predictions generated by such a model may more accurately reflect the perception of item desirability among the general population, in contrast to studies employing smaller samples of human judges. The findings of this study provide an initial, albeit modest, indication supporting this hypothesis, as the data demonstrated that machine-rated item desirability exhibited a similar alignment with the judgments of both laypeople and psychology students.

In conclusion, this study represents an important step forward in the use of advanced natural language processing techniques to automatically obtain estimates of item desirability. With further research and refinement, this method has the potential to transform the way researchers and practitioners measure social desirability bias.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.paid.2023.112307.

CRediT authorship contribution statement

Björn E. Hommel: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

Data availability

Materials, data, and code for the present study are available through the Open Science Framework: https://osf.io/67mkz/.

Acknowledgments

I want to thank Lucy Erber and Paulina Machado Costa for their support in reviewing the literature for this research. I am also thankful to Dr. Hannes Zacher for providing valuable feedback on this manuscript.

References

Andersen, H., & Mayerl, J. (2019). Responding to socially desirable and undesirable topics: Different types of response behaviour? *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology (Mda), Data.*. https://doi.org/10.12758/MDA.2018.06

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology, 9*(3), 272–279. https://doi.org/10.1037/h0025907

Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2022). XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond (arXiv:2104.12250). arXiv. doi:10.48550/arXiv.2104.12250.

Bochner, S., & Van Zyl, T. (1985). Desirability ratings of 110 personality-trait words. *The Journal of Social Psychology, 125*(4), 459–465. https://doi.org/10.1080/00224545.1985.9713524

Britz, S., Gauggel, S., & Mainz, V. (2019). The Aachen list of trait words. *Journal of Psycholinguistic Research, 48*(5), 1111–1132. https://doi.org/10.1007/s10936-019-09649-8

Britz, S., Rader, L., Gauggel, S., & Mainz, V. (2022). An English list of trait words including valence, social desirability, and observability ratings. *Behavior Research Methods*, 1–18.

Chandler, J. (2018). Likeableness and meaningfulness ratings of 555 (+487) person-descriptive words. *Journal of Research in Personality, 72*, 50–57. https://doi.org/10.1016/j.jrp.2016.07.005

Converse, P. D., Pathak, J., Quist, J., Merbedone, M., Gotlib, T., & Kostic, E. (2010). Statement desirability ratings in forced-choice personality measure development: Implications for reducing score inflation and providing trait-level information. *Human Performance, 23*(4), 323–342. https://doi.org/10.1080/08959285.2010.501047

Dumas, J. E., Johnson, M., & Lynch, A. M. (2002). Likableness, familiarity, and frequency of 844 person-descriptive words. *Personality and Individual Differences, 32*(3), 523–531. https://doi.org/10.1016/S0191-8869(01)00054-X

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.* Dryden Press.

Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., Glorioso, M., & Hack, E. (2023). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0001082

Filter, J. (2018, December 6). *Clean-text/clean.py at main · jfilter/clean-text*. GitHub. https://github.com/jfilter/clean-text.

Fyffe, S., Lee, P., & Kaplan, S. (2023). "Transforming" personality scale development: Illustrating the potential of state-of-the-art natural language processing. *Organizational Research Methods, 109442812311557*. https://doi.org/10.1177/10944281231155771

Gal, Y., & Ghahramani, Z. (2016). *Dropout as a Bayesian approximation: Representing model uncertainty in deep learning* (arXiv:1506.02142). arXiv. doi:10.48550/arXiv.1506.02142.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*(1), 84–96.

Götz, F. M., Maertens, R., Loomba, S., & van der Linden, S. (2023). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*. https://doi.org/10.1037/met0000540

Hampson, S. E., Goldberg, L. R., & John, O. P. (1987). Category-breadth and social-desirability values for 573 personality terms. *European Journal of Personality, 1*(4), 241–258. https://doi.org/10.1002/per.2410010405

Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-based deep neural language modeling for construct-specific automatic item generation. *Psychometrika, 87*(2), 749–772. https://doi.org/10.1007/s11336-021-09823-9

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics, 35*(1), 73–101. https://doi.org/10.1214/aoms/1177703732

Hughes, A. W., Dunlop, P. D., Holtrop, D., & Wee, S. (2021). Spotting the "ideal" personality response: Effects of item matching in forced choice measures for personnel selection. *Journal of Personnel Psychology, 20*(1), 17–26. https://doi.org/10.1027/1866-5888/a000267

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity, 47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9

Lee, J., Tang, R., & Lin, J. (2019). What would Elsa do? Freezing layers during transformer fine-tuning. *ArXiv, 1911*, Article 03090. http://arxiv.org/abs/1911.03090.

Lee, P., Fyffe, S., Son, M., Jia, Z., & Yao, Z. (2022). A paradigm shift from "human writing" to "machine generation" in personality test development: An application of state-of-the-art natural language processing. *Journal of Business and Psychology*. https://doi.org/10.1007/s10869-022-09864-6

Leising, D., Vogel, D., Waller, V., & Zimmermann, J. (2021). Correlations between person-descriptive items are predictable from the product of their mid-point-centered social desirability values. *European Journal of Personality, 35*(5), 667–689.

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:1907.11692). arXiv. doi:10.48550/arXiv.1907.11692.

McIntyre, M. M. (2022). Judging what others enjoy: Desirability and observability of interests. *Journal of Career Assessment, 30*(3), 557–572. https://doi.org/10.1177/10690727211055862

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15*(3), 263–280. https://doi.org/10.1002/ejsp.2420150303

Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner, & J. S. Wiggins (Eds.), *Personality assessment via questionnaires* (pp. 143–165). Berlin Heidelberg: Springer. https://doi.org/10.1007/978-3-642-70751-3_8.

Pavlov, G., Shi, D., Maydeu-Olivares, A., & Fairchild, A. (2021). Item desirability matching in forced-choice test construction. *Personality and Individual Differences, 183*, Article 111114. https://doi.org/10.1016/j.paid.2021.111114

Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *ArXiv:2004.09813 [Cs]*. http://arxiv.org/abs/2004.09813.

Rounds, J., Su, R., Lewis, P., & Rivkin, D. (2010). *O* NET interest profiler short form psychometric characteristics: Summary*. Raleigh, NC: National Center for O* NET Development.

Ryan, A. M., Bradburn, J., Bhatia, S., Beals, E., Boyce, A. S., Martin, N., & Conway, J. (2021). In the eye of the beholder: Considering culture in assessing the social desirability of personality. *Journal of Applied Psychology, 106*(3), 452.

Schönbach, P. (1972). Likableness ratings of 100 German personality-trait words corresponding to a subset of Anderson's 555 trait words. *European Journal of Social Psychology, 2*(3), 327–333. https://doi.org/10.1002/ejsp.2420020309

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Urban, C. J., & Gates, K. M. (2021). Deep learning: A primer for psychologists. *Psychological Methods*. https://doi.org/10.1037/met0000374

van Genugten, R., & Schacter, D. L. (2022). *Automated scoring of the autobiographical interview with natural language processing. PsyArXiv. January, 23*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008. https://arxiv.org/abs/1706.03762.

Wessels, N. M., Zimmermann, J., & Leising, D. (2021). Who knows best what the next year will hold for you? The validity of direct and personality–based predictions of future life experiences across different perceivers. *European Journal of Personality, 35*(3), 315–339. https://doi.org/10.1002/per.2293

Wetzel, E., Frick, S., & Brown, A. (2021). Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking. *Psychological Assessment, 33*(2), 156.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). https://doi.org/10.18653/v1/2020.emnlp-demos.6

Wood, J. K., Anglim, J., & Horwood, S. (2022). A less evaluative measure of Big Five personality: Comparison of structure and criterion validity. *European Journal of Personality, 36*(5), 809–824. https://doi.org/10.1177/08902070211012920