Empirical Study

PERSONALITY Science

Personality Science
Volume 6: 1–13

DOI: 10.1177/27000710251391608
journals.sagepub.com/home/ppp
Sage

Properties of Person Descriptors in the Natural German Language: A Preregistered Replication and Extension

Moritz Leistner[1], Björn E. Hommel[2], Leon P. Wendt[3] and Daniel Leising[4]

Abstract

Understanding language is crucial to psychological research, as it is the basis for most psychological measurements. Building on previous work, we conducted a preregistered replication study, analyzing 876 person descriptors generated by 187 participants using a free response format. These person descriptors were rated for social desirability, observability, importance, abstractness, base rate, and stability by approximately 15 human raters each (n = 456). Key findings were replicated, including a bimodal distribution of social desirability, a greater number of negative vs. positive person descriptors, and a U-shaped relationship between importance and social desirability. Furthermore, human ratings of social desirability could be closely approximated using a fine-tuned encoder model and GPT-4o. However, GPT-4o's performance in approximating human ratings of the other person descriptor properties showed some deviations. This suggests that, despite showing potential for more economical data collection, there is significant room for improvement in using AI-applications for emulating human ratings of natural language person descriptors.

Keywords

natural language processing, psycholinguistics, social desirability, large-language model, transformer-models

Received: 2 July 2024; revised: 6 October 2025; accepted: 13 October 2025

Introduction

The natural language is the basis for most psychometric measurements, including measurements of personality. Given that most self- and other-ratings of personality use items phrased in natural language terms, it is important to develop a systematic understanding of the properties of those terms. Without such an understanding, there is substantial risk of misinterpreting research findings. Correlations between items may be interpreted as "substantive" (i.e., reflecting associations between the behaviors that they refer to) when in fact they may be rooted in other influences (e.g., shared evaluation; Leising et al., 2025). Generally speaking, different term properties may shape judgments by way of meaningful interactions with properties of perceivers, targets, and perceiver-target dyads.

In the present paper, we draw a representative sample of personality-descriptive adjectives (hereafter referred to as *descriptors*) and jointly investigate six key properties that distinguish these descriptors from one another (Leising et al., 2014; Wessels et al., 2025): Social desirability, observability, base rate, abstractness, stability, and importance. We determine how reliable properties can be rated by human participants, how the properties are distributed, and how ratings statistically relate to one another. In doing so, we attempt a near-identical, preregistered replication of previous findings reported by Leising et al. (2012, 2014). By publishing our data, we also aim to

[1]Otto-Friedrich University Bamberg, Bamberg, Germany
[2]University of Leipzig, Leipzig, Germany
[3]University of Kassel, Kassel, Germany
[4]TU Dresden, Dresden, Germany

Corresponding Author:
Moritz Leistner, Chair of Personality Psychology and Psychological Assessment, Otto-Friedrich-University Bamberg.
Email: moritz.leistner@uni-bamberg.de

enable more generalizable and replicable studies of person judgments using the natural language. In addition, we derive synthetic ratings of the same sample of descriptors by Large Language Models, and determine their comparability to human ratings. The use of synthetic ratings may pave a way for greater research efficiency in the future.

Descriptor Properties

Social desirability denotes the extent to which the use of a descriptor to describe a person sheds a positive light on that person (Edwards, 1953). Notably, the same behaviors can be portrayed positively (e.g., "meticulous") or negatively (e.g., "obsessive") by using descriptors varying in social desirability (Bäckström et al., 2009; John & Robins, 1994; Leising et al., 2015; Nederhof, 1985; Peabody, 1967; Wood et al., 2021).

Observability pertains to how visible or detectable a trait is to an observer (Funder & Dobroth, 1987), that is, how easily it can be inferred from observed behavior.

Abstractness pertains to the breadth of different behaviors that are relevant to describing persons with a descriptor (Burgoon et al., 2013; Della Rosa et al., 2010; Hampson et al., 1986; John et al., 1994; Mõttus et al., 2017).

Stability concerns the extent to which the use of a descriptor implies a claim that the referenced behaviors are stable over time (Bleidorn, 2012; Wood & Wortman, 2011).

Importance reflects the perceived relevance of knowing that certain traits are being attributed to someone (Cottrell et al., 2007; Williams et al., 1998; Wood, 2015).

Base Rate is the proportion of persons to which a descriptor is assumed to apply (Funder & Dobroth, 1987).

It needs to be acknowledged that there is some conceptual ambiguity associated with some of these descriptor properties. Further, while these properties are conceptually distinct, previous research has shown that, empirically, there is some overlap between them. For example, positive descriptors seem to be more abstract than negative ones, an example of *valence asymmetry* (Baumeister et al., 2001; Unkelbach et al., 2008).

Replication Targets

We attempted to replicate some key findings from a previous study by Leising et al. (2012, 2014), and expected the following: (a) Interrater reliability (ICC(2,k)) for all six properties exceeding *0.80*. (b) a predominance of negative over positive descriptors of approximately 60%, (c) social desirability ratings of negative descriptors that are more extreme (i.e., evaluative) than those of positive descriptors, (d) a pattern of correlations among the six descriptor properties similar to the one obtained by Leising et al. (2014), for both zero-order and partial correlations ($r >$ *.80)*, (e) social desirability ratings exhibiting a bimodal distribution, but unimodal distributions for all other descriptor properties. These expectations were pre-registered before data collection.

Synthetic Ratings

Recent advances in the field of natural language processing have yielded large language models (LLMs) which increasingly demonstrate value for the social and behavioral sciences. For example, research by Hommel (2023) indicated that transformer models can be trained to predict how survey respondents perceive item desirability (r = .80). Our study aimed to replicate this finding, focusing on descriptors not previously rated or used in the training process, to validate synthetic ratings against human social desirability ratings. However, since Hommel's transformer model could not be used to rate descriptor properties *other* than social desirability, we also used OpenAI's GPT-4o (OpenAI, 2024), a general-purpose LLM, to try to approximate human ratings of person descriptors on all six properties. Substituting human raters with synthetic estimates has the potential to drastically increase the cost-effectiveness and feasibility of collecting such ratings, as recruiting human raters is expensive and labor-intensive.

Methods

The authors declare that they have no conflicts of interest. The study consisted of three preregistered phases: descriptor generation, descriptor rating, and comparison of synthetic with human ratings. All funding came from internal resources. Hypotheses and methods for each phase were preregistered (2022-09-25, 2022-10-28, 2023-11-13) before data collection and are publicly available. The analysis plan was not preregistered. All materials, data, and annotated R scripts are publicly available. AI was used to refine the writing.

Participants

A total of 643 participants, fluent in German, were recruited, with 187 in the first phase and 456 in the second. Participants varied in age (M = 39.9, SD = 14.7) and were predominantly female (79.7% in the first, 74.6% in the second phase). The first sample showed a tendency to contain a disproportionate amount of participants with high education (29.4% had no or low secondary education, 43.9% had higher secondary education and 23% held a university degree), and this tendency was even more pronounced in the second sample (7.7% had no or low secondary education, 54.4% had higher secondary education and 36% held a university degree). Descriptive statistics of sociodemographic measures of both samples can be found in supplemental files in the OSF folder.

Recruitment methods varied, with university and social media ads for the first phase and mainly Meta (Facebook) for the second. Monetary rewards of varying sizes were offered across phases (Sweepstakes for the first phase and €10 per participant for the second). The target sample size for the first phase was 150 participants, based on prior studies (Leising et al., 2012, 2014). The target sample size for the second phase was 450 raters, which was determined through an a priori power analysis to achieve an ICC(2,k) of .80 ($\alpha = .05$). However, we planned to collect data from 500 individuals to account for exclusions due to careless responding (Arifin, 2023; Bonett, 2002; Walter et al., 1998).

Materials

The first phase was used to generate a comprehensive and representative sample of descriptors, as we employed the methodology established by Leising et al. (2012, 2014). Participants were instructed to describe themselves and people they knew and liked to varying degrees. Specifically, they were asked to think of someone they knew well and liked (e.g., a family member), someone they knew well but did not like (e.g., a disliked colleague), someone they did not know well but liked (e.g., a celebrity), and someone they neither knew well nor liked (e.g., a disliked boss). For each category, participants were requested to provide between 3 and 10 descriptors.

In the second phase, each participant rated 175 descriptors on one descriptor property (e.g., social desirability) assigned to them. Raters were provided with a definition of the descriptor property and used a percentile scale (from 0 to 100) to indicate how strongly the descriptor property applied to each descriptor. The number of descriptors was chosen to allow each set of ratings to be completed in a single sitting.

The verbal anchors used for these ratings can be found in Table 1. Complete details of the instructions and all materials used in the study can be found in the preregistration (in the OSF respository).

Both phases were conducted online. Participants provided socio-demographic data about themselves and descriptions of targets, which were then rated in the second phase along the properties. Additionally, participants evaluated a number of public figures. This data was not analyzed for the present study.

Analysis of Synthetic Ratings

Synthetic ratings were generated using a transformer model trained to predict human judgments of social desirability in survey items (Hommel, 2023; Hussain et al., 2023). This model built on a sentiment classifier originally designed to predict categorical valence (Barbieri et al., 2022). Given the strong link between valence and social

Table 1. Verbal Rating Scale Anchors for Phase Two Rating Instructions

Property	Minimum anchor	Maximum anchor
Observability	Extremely difficult to observe (“extrem schlecht beobachtbar”)	Extremely easy to observe (“extrem gut beobachtbar”)
Abstractness	Extremely concrete (“extrem konkret”)	Extremely abstract (“extrem abstrakt”)
Stability	Extremely unstable (“extrem instabil”)	Extremely stable (“extrem stabil”)
Importance	Extremely unimportant (“extrem unwichtig”)	Extremely important (“extrem wichtig”)
Base rate	Extremely rarely applicable (“extrem selten anwendbar”)	Extremely frequently applicable (“extrem häufig anwendbar”)
Social desirability	Extremely negative (“extrem negativ”)	Extremely positive (“extrem positiv”)

Note. Participants rated all descriptors on a 0 to 100 scale. The minimum anchor represented the value for 0, the maximum anchor represented the value for 100.

desirability (e.g., Britz et al., 2019, 2022), the model successfully adapted to predict z-standardized desirability scores having used data from 14 independent studies. For previously unseen items, synthetic estimates correlated strongly with human ratings (r = .80), which clearly outperformed the original sentiment model (r = .66; Hommel, 2023). However, being initially trained as a sentiment classifier could have still limited the alignment with human ratings due to some conceptual differences between sentiment and desirability.

Determining the "neutral point" for synthetic ratings presented a bit of a challenge. Consistent with Leising et al. (2014) and our own findings (see Results), participants, who were given a 0 to 100 scale tended to produce bimodal distributions of social desirability, with more negative than positive ratings. As a result, the empirical mean of such ratings tended to lie slightly below the numerical midpoint of 50. The LLM, however, was trained on z-standardized human ratings, where the mean (M = 0) corresponded to this slightly negative empirical value. This means that, unlike on the 0 to 100 scale where 50 was defined as neutral by convention, the LLM's "neutral" point would have been set to a value that is already somewhat negative. To correct this misalignment, we calculated an adjusted mean, defined by the average synthetic rating for the 30 descriptors rated closest to neutral (i.e., 50 on the 0 to 100 scale) by humans. This adjustment aligned the LLM's point of neutrality more closely with that of human perception.

We also used OpenAI's GPT-4o (OpenAI, 2024) application programming interface (API), a general purpose LLM, operated through R Studio to obtain synthetic ratings for all properties, not just social desirability. Unlike ChatGPT, the API processed each request independently, with no memory of prior interactions with the user. Each descriptor produced by human participants in the first phase, and rated by human participants in the second phase, was also rated by GPT-4o. GPT-4o received the same instructions and rating scale as human raters, without any role-specific prompting or changes to hyperparameters. These ratings were obtained using default, non-deterministic settings, which meant that the LLM did not typically assign exactly the same rating to the same descriptor in repeated trials. To account for this stochastic variability, each descriptor was rated 15 times, and the ratings were then aggregated. Reliability across trials is reported below. Given the highly exploratory nature of these analyses, we did not preregister them.

Data Quality and Careless Responding

The study achieved its target sample size in both phases. In phase one (descriptor generation), 187 participants completed the questionnaire. In phase two (descriptor ratings), 587 participants completed the questionnaire. No careless-response checks were needed for phase one due to its open-ended format. Phase two offered a €10 incentive and used checks for careless responses, including any failure to select "can't answer" when presented with fake descriptors or the excessive use of the "can't answer" option for real descriptors (>15%). Instructed response items were included as well, but not used for exclusion due to technical issues on mobile devices. After 153 careless respondents were excluded, a total of 456 valid cases were retained and used for analysis. We note that the high exclusion rate (26%) may have partly been due to the decision to remove participants with excessive use of the "can't answer" option, given that such response patterns may have sometimes reflected genuine uncertainty rather than careless responding.

Results

Data Preparation and Analysis Plan

In accordance with the second preregistration, descriptors were excluded if they described physical attributes, nouns or verbs without an equivalent adjective form, references to real human beings (e.g., "Albert Einstein"), or words outside the German language or with unclear meaning, as defined by the dictionary (Duden, 2020). In several instances, we modified the descriptors (as provided by participants) in order to align them with the above criteria. If possible, words were adapted to their adjective forms, rare variants were replaced with more common ones (e.g., "devotedly" → "devoted"), and typos were corrected in order to lemmatize the words. Additionally, for descriptors not typically found in the Duden's German language corpus (e.g., "unsuperficial") that were still grammatically valid and understandable, exceptions were made to retain them. The final list was then compared to the list of Leising et al. (2014).

As we expected more negative than positive descriptors, we split descriptors into two categories based on whether they fell above or below the scale midpoint of 50 (on a 0–100 scale). A two-sided binomial test was used to compare the proportion of positive vs. negative descriptors. An independent samples t-test then assessed whether negative descriptors were more extreme than positive ones. To compare the correlational pattern between properties with the one found in Leising et al. (2014), we computed partial correlations and 95% confidence intervals, and then compared zero-order correlations using Pearson's r. To examine the expected U-shaped relationship between importance and social desirability, we analyzed the correlation between importance and squared z-standardized social desirability, visualized with a scatterplot.

Replication Attempt

The first phase generated 4,369 person descriptions, categorized as self-descriptions (25.6%), liked acquaintances (21.7%), disliked acquaintances (19.8%), liked but less known individuals (17.1%), and neither known nor liked individuals (15.8%). Among the most common descriptors, only 4 of the top 15 were negative (*'egotistical', 'arrogant', 'dishonest' & 'pretentious'*). From 1,285 distinct descriptors identified, only 876 remained after following the steps defined earlier. Of those descriptors, 413 (66.2%) were shared with the 624 descriptors retained by Leising et al. (2012), which showed a significant overlap that was still short of the predicted 80%. Both the initial list with frequencies and the refined list of descriptors are available in the OSF folder.

For the second phase, each of the 456 participants rated around 175 descriptors. On average, each descriptor received ratings from 15.2 raters on each property. Table 2 presents descriptive statistics and distributions of the rated properties.

ICC(2,k) values quantifying inter-rater reliability are presented in Table 2 and in the supplemental files. In the preregistration, based on the results of Leising et al. (2014), we had established .80 as the expected lower threshold for these values. Only ratings of social desirability (.98) surpassed this expectation, whereas ratings of observability (.78), importance (.75), base rate (.73) and stability (.79) narrowly missed it. Inter-rater reliability for ratings of descriptor abstractness was substantially lower (.57).

As most of these values were lower than in the previous study by Leising et al. (2014), it was clear that there was more variance in the way participants rated descriptor properties. On the one hand, we speculate that having used an online questionnaire in the present study may have incurred a certain proportion of systematic error that we were unable to detect. However, we also consider it possible that the increased variance might reflect less shared idiosyncratic understanding of descriptor properties and may just be a reflection of a more diverse sample (in terms of language use) than the previous studies.

Further, it could be shown that social desirability, unlike the other properties, was bimodally distributed, which aligned with preliminary findings (Figure 1). A two-tailed binomial test revealed that 56.3% of descriptors were negative, significantly exceeding positive descriptors ($p < .001$). We used Cohen's h as an estimate for effect size and found h = 0.25, which indicated a small effect (Cohen, 1988) and confirmed the hypothesis. Furthermore, negative descriptors showed a greater average negativity than the positive descriptors showed average positivity (Welch t-test: $t(876) = 5.44$, $p < .001$). This was interpreted as a small effect (Cohen's d = 0.37; Cohen, 1988), indicating that positive descriptors were less evaluative than negative descriptors, thereby supporting the initial prediction. While we predicted this, the result was not trivial,

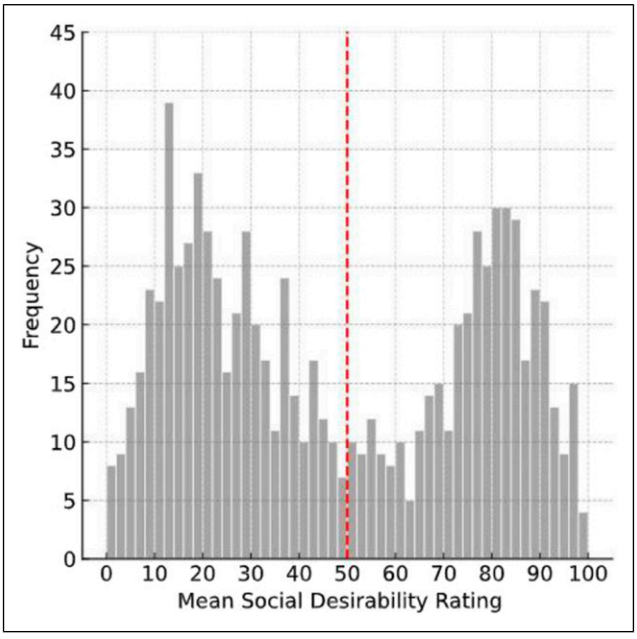

Figure 1. Frequency Distribution of the Social Desirability Ratings of the 876 Descriptors. The Rating Scale Ranged From 0 (Low) to 100 (High)

Table 2. Human Ratings of Person Descriptors: Interrater-Reliability, Means, Standard Deviations and High and Low Average Ratings

Property	Reliability	*M*	*S*	Descriptors with the highest average ratings	Descriptors with the lowest average ratings
Observability	.78	58.6	12.4	Well-groomed, talkative, loud	Bisexual, religious, prone to addiction
Abstractness	.57	43.5	10.8	Difficult, intense, dumb	Bisexual, despotic, misogynous
Stability	.79	57.4	13.1	Positive, family-oriented, gay	Whacked, confused, offended
Importance	.75	52.4	13.0	Honest, respectful, reliable	Gay, unique, unathletic
Base rate	.73	48.3	13.0	Humane, nice, confident	Bigoted, war-fanatic, undignified
Social desirability	.98	45.3	27.6	Trustworthy, reliable, respectful	Homophobic, violent, power-abusing

Note. 876 descriptors were rated by 456 participants. M = mean, *S* = standard deviation, Reliabilities are represented by ICC(2,k) values.

considering that the instruction required to describe two liked targets and the self (which were generally also assumed to be liked), compared to only two targets that were unliked. Accordingly, participants also provided the highest average amount of descriptors for the self and the known and liked target (25.6%/21.7%). Just based on probability, the opposite hypothesis (i.e., more positive descriptors) could have been assumed, underscoring the significance of this finding.

The correlations of the properties were analyzed as zero-order and partial correlations (i.e., the latter controlling for the other properties). Furthermore, the correlation pattern of the new sample was compared with the one from Leising et al. (2014). Table 3 presents the correlations of the correlation patterns of both samples. Overall, the zero-order correlations in the two studies correlated at r(*13*) = .66 with one another, p = .007, whereas the partial correlations correlated at r(*13*) = .80, $p < .001$, respectively.

Additionally, we investigated the relationship between importance and social desirability (see Figure 2). The basic zero-order correlation between social desirability and importance was r(*874*) = .34, $p < .001$, with a partial correlation of *r(874)* = *.04*, $p < .001$, after having been adjusted for the other properties. Notably, the correlation for squared social desirability (after z-transformation) and importance was significantly stronger, with a Pearson correlation of r(*874*) = .55 ($p <$ .001, 95% CI [.50, .59]), as shown in Figure 2, and comparable to Leising et al. (2014) that found a correlation of r(*621*) = .63, $p < .001$. This highlights a pronounced relationship between evaluativeness and perceived importance.

Statistical Analysis of Synthetic Ratings

Next, we investigated the performance of Hommels' encoder model by comparing its synthetic ratings of social desirability to the human ratings. At first, we exclusively considered descriptors that did not overlap with the models' original training dataset, as we were interested in generating an estimate for the models' performance with unknown descriptors. In this context, the correlation of synthetic estimates and human ratings amounted to r(*503*) = .65, which did not meet the expected magnitude of r = .80. When we investigated descriptors that the encoder model was trained on, a significant and high correlation was observed, r(*367*) = .92. While the ratings of new person descriptors followed a bimodal distribution, in accordance with hypothesis 2 (Figure 3), in opposition to empirical findings of human ratings, only 51.2% of the descriptors were negative, despite having used the adjusted mean. Further, unlike with human ratings, positive descriptors (*M* = *.88, σ* = *.37*) were more evaluative than negative descriptors (*M* = *.75, σ* = *.36*). This can be

Table 3. Zero-Order and Partial Correlations Between the Properties in the 2014 Sample and in the Current Sample

	Social desirability				Abstractness				Stability				Importance				Base rate			
	zero-order		partial		zero-order		partial		zero-order		partial		zero-order		partial		zero-order		partial	
Property	*2014*	*2024*	*2014*	*2024*	*2014*	*2024*	*2014*	*2024*	*2014*	*2024*	*2014*	*2024*	*2014*	*2024*	*2014*	*2024*	*2014*	*2024*	*2014*	*2024*
Observability	.04	.15*	.00	.07	.00	−.16*	−.01	−.16*	.04	.11*	.03	−.06	−.09*	.14*	−.10*	.01	.05	.34*	.04	.28*
Social desirability					.24*	.22*	.23*	.40*	.47*	.59*	.41*	.53*	.07	.35*	−.04	.04	.56*	.46*	.47*	.32*
Abstractness									−.06	−.12*	−.20*	−.27*	.00	−.11*	.00	−.08*	.18*	−.05	.07	−.08*
Stability													.13*	.47*	.11*	.30*	.31*	.33*	.06	−.02
Importance																	.13*	.36*	.10*	.21*

Note. Zero-order correlations are Pearson product-moment correlations. To determine the partial correlations, the R package ppcor was used. *p < .05. Here, the four term properties not being correlated with one another were partialled out of the two that were correlated with one another.

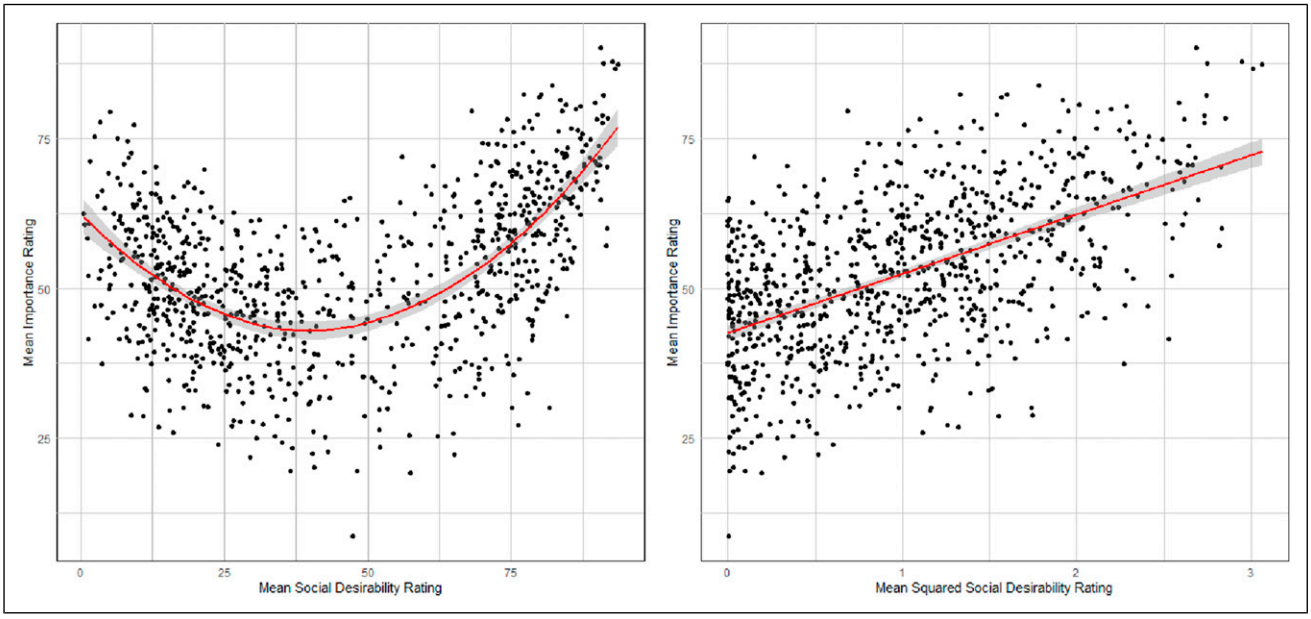

Figure 2. Scatter Plot of Importance and Social Desirability

interpreted as a significant ($t(499) = -3.91$, $p < .001$) and small effect ($d = .35$).

Table 4 displays partial correlations between descriptor properties using different desirability ratings (2014 vs. Encoder-Model vs. the current sample). Lastly, we tried to replicate the relationship between importance and social desirability ratings, found in both our empirical data sets. We found it to be less pronounced, with a zero-order correlation of $r(503) = .32$ between the squared encoder-model social desirability ratings and human importance ratings.

Following this, we also examined synthetic ratings of GPT-4o. Since the training data of GPT-4o was not (and most likely won't ever be) publicly available, we were unable to determine which descriptors were included in its training, and we had assumed that, in fact, all were. Therefore, we used the full sample of descriptors to compute correlations between human and synthetic ratings. When we investigated the ability to generate social desirability ratings similar to empirical human ratings, GPT-4o outperformed the encoder-model's results for trained-on person descriptors slightly: We found a

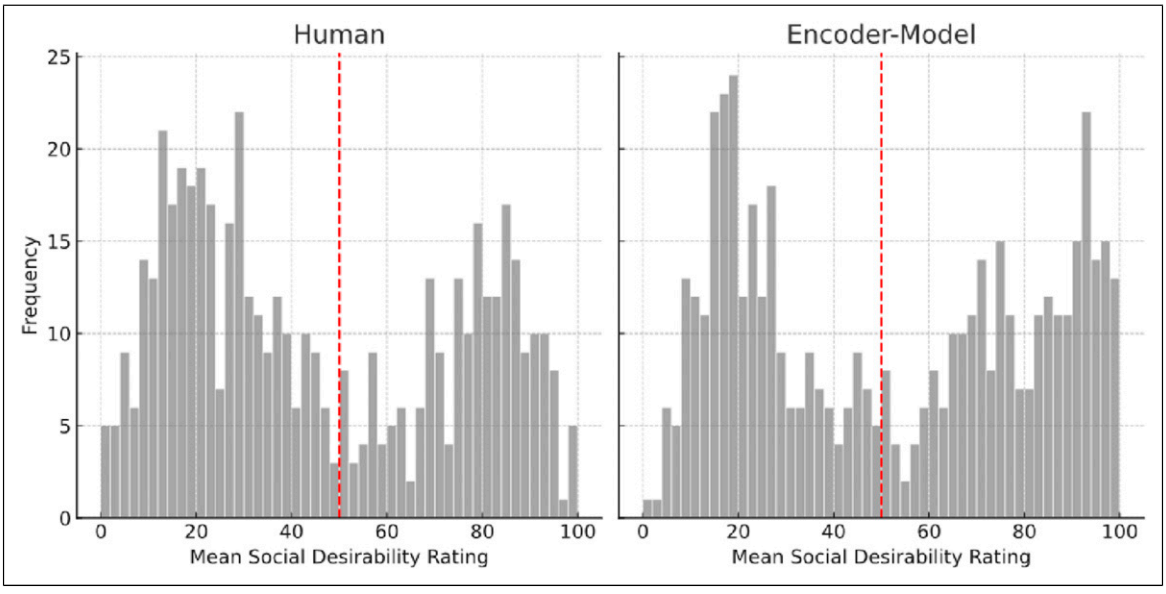

Figure 3. Frequency Distributions of Social Desirability Ratings for New Person Descriptors

Table 4. Partial Correlations of All Data Sets of Social Desirability With Other Properties

	Observability			Importance			Stability			Abstractness			Base rate		
	2014	*EM*	*CHS*	*2014*	*EM*	*CHS*	*2014*	*EM*	*CHS*	*2014*	*EM*	*CHS*	*2014*	*EM*	*CHS*
Social desirability	.00	.08*	.07	−.04	.10*	.04	.41*	.42*	.53*	.23*	.18*	.40*	.47*	.21*	.32*

Note. To determine the partial correlations, the R package ppcor was used. *$p < .05$. Here, the four term properties not being correlated with one another were partialled out of the two that were correlated with one another. EM = encoder model, CHS = current human sample.

correlation of r(*872*) = .95 ($p < .001$; compared to r(*367*) = .92), following a bimodal distribution, as can be seen in Figure 4. GPT-4o ratings also, different to the encoder-model, replicated the empirical findings of more negative than positive descriptors (56.5% negative) and those negative descriptors to be more evaluative (t(*854*) = 2.71, $p = .007$, d = .18). While GPT4o ratings seemed to closely mimic human ratings for social desirability, we however found that GPT4o overall produced more extreme ratings than human judges.

Additionally, we investigated ratings on other properties than social desirability. While the human ratings, besides the bimodal distribution of social desirability, all followed an unimodal distribution, this was not the case for the GPT-4o ratings. Histograms can be found in the supplement. Importance and Base Rate followed a somewhat bimodal distribution, while Abstractness followed an unimodal, but left-skewed distribution. On the other hand, the distribution of Importance and Observability could have even been considered trimodal.

Further, we investigated how well the correlative structure of GPT-4o ratings compared to that of human ratings. Again using ICC(2,k), we found almost perfect reliability for GPT-4o ratings of all six descriptor properties (Table 5).

Table 6 shows the pattern of zero-order correlations among the six descriptor properties, within human ratings, within synthetic ratings, and between the two types of ratings. Correlations in parentheses were disattenuated for inter-rater unreliability. The pattern of correlations among descriptor properties converged at only r(*13*) = .37 ($p = .18$) between the two types of ratings. However, the main reason for this relatively low and non-significant correlation was clearly abstractness. Without it, the correlation was r(*8*) = .71 ($p = .02$).

Some descriptor properties seemed particularly difficult to disentangle, as evident from their high (disattenuated) correlations: Synthetic ratings of base rate correlated at r(*872*) = .78 with human ratings of base rate, but also at r(*872*) = .71 with human ratings of social desirability. Human ratings of stability correlated at r(*872*) = .67 with synthetic ratings of stability, but also at r(*872*) = .71 with synthetic and human ratings of desirability, respectively.

We further explored which synthetic ratings differed the most from human ratings, as well as which descriptors got rated the highest and lowest by GPT-4o (Table 5). A complete list can be found in the supplemental files. GPT-4o sometimes missed meanings humans would have intuitively grasped and rated descriptors more literally (e.g., "unstable" rated as unstable, z = −2.87).

Lastly, we investigated if we could replicate the empirical finding of an U-shaped correlation between social desirability and importance (Leising et al., 2014), surprisingly having found that this effect was even more pronounced in GPT-4o ratings. This was likely due to the fact that GPT-4o provided more extreme ratings in general. Scatter Plots for this interaction can be found in the supplemental material in the OSF-folder.

Discussion

This study had three goals: First, to generate a large, representative sample of person descriptors from the natural (German) language and have them rated on six key properties. Second, to test the replicability of key findings from Leising et al. (2012, 2014), using strict preregistration. Third, to explore the extent to which synthetic ratings can substitute human ratings, potentially offering a more cost- and time-efficient alternative.

Sampling and Rating of Descriptors

Having used the same methodology as Leising et al. (2012), we collected a new, large sample of person descriptors ($n = 876$). Since a substantial number of perceivers independently generated these via free-response descriptions of real targets, we had no reason to doubt that this sample could be fairly representative of the natural person-descriptive lexicon despite the use of a highly educated sample.

While 66.2% of the new descriptor sample did overlap with the sample collected by Leising et al. (2014), there also was substantial non-redundancy (211 unique to the first study, 461 unique to the second). This suggests that the person-descriptive lexicon is extremely differentiated and does contain hundreds of descriptors that are almost never used. Although inter-rater reliabilities slightly missed the preregistered benchmarks, they were still satisfactory, and raters were able to distinguish the six descriptor properties from one another. We thus feel

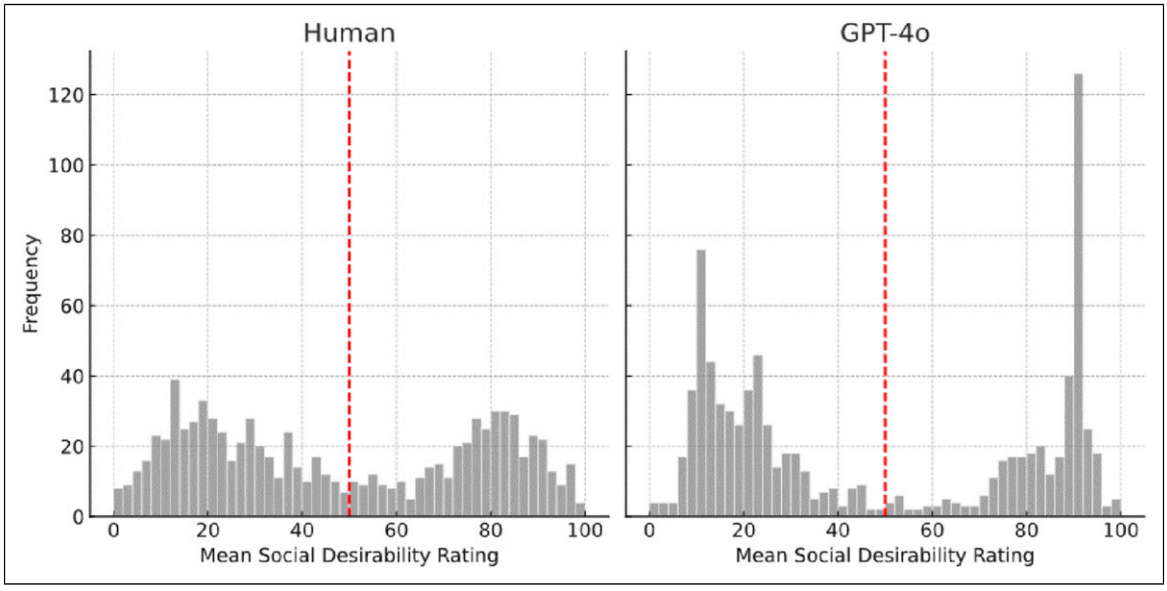

Figure 4. Frequency Distribution of GPT-4o Social Desirability Ratings Compared With Human Ratings

confident recommending this dataset (openly accessible at OSF) to researchers interested in using it for their own work.

Replicability of Key Findings

We were able to replicate several previous findings using strict preregistration. We confirmed the bimodal distribution of social desirability ratings, the unimodal distributions of all other property ratings, a curvilinear association between social desirability and importance (r(*874*) = .55), a greater proportion of negative as compared to positive descriptors (56%, h = .25), and a greater evaluativeness of negative descriptors (d = .37). The latter two findings suggest a general kind of valence asymmetry (Baumeister et al., 2001; Unkelbach et al., 2008), with negative person descriptions having been more differentiated and more emotionally intense than positive ones, on average.

Furthermore, we found that the pattern of partial correlations among the six descriptor properties closely replicated the one reported in Leising et al. (2014, Table 3), *r(13)* = *.80*. As in the previous study, more positive descriptors were also rated as being more traitlike, more abstract, and applicable to a larger number of people (see Table 3). This finding again pointed to a valence asymmetry, with positive person descriptions being broader (i.e., less differentiated). Given the fact that the two studies were conducted ten years apart, with no overlap in participant samples, and having used a strict replication strategy, we are quite confident that these effects may now be viewed as being relatively robust.

Investigating Synthetic Ratings

This study also assessed how well two Large Language Models (LLMs) were able to emulate human judgments of descriptor properties. Both LLMs performed well on trained descriptors (.92/.95) for ratings of social desirability. However, the encoder model did less well on new descriptors (.65), possibly due to occasionally having conflated sentiment with desirability (Hommel, 2023).

Table 5. GPT-4o Ratings of Person Descriptors: Interrater-Reliability, Means, Standard Deviations and High and Low Average Ratings

Property	Reliability	*M*	*S*	Descriptors with the highest average ratings	Descriptors with the lowest average ratings
Observability	.98	52.8	17.4	Feminine, masculine, ungroomed	Full of integrity, bisexual, pain-sensitive
Abstractness	.97	62.2	16.3	National-socialist (Nazi), positive, human	Hurt, stubborn, male
Stability	.99	63.8	17.1	Feminine, masculine, gay	Unique, unstable, variable
Importance	.99	65.3	17.1	National-socialist (Nazi), racist, violent	Full of integrity, wishy-washy, cuddly
Base rate	.99	42.1	19.4	Human, pain-sensitive, teachable	Autistic, national-socialist (Nazi), full of integrity
Social desirability	1.00	45.1	31.9	Role-model, kind-hearted, stunning	National-socialist (Nazi), racist, misogynistic

Note. 876 descriptors were rated by 15 API requests per descriptor. M = mean, *S* = standard deviation, Reliabilities are represented by ICC(2,k) values.

Table 6. Zero-Order and Disattenuated Zero-Order Correlations Among Descriptor Properties

Variable		Observability Human	Social desirability Human	Abstractness Human	Stability Human	Importance Human	Base rate Human	Observability GPT-4o	Social desirability GPT-4o	Abstractness GPT-4o	Stability GPT-4o	Importance GPT-4o
Social desirability	Human	.15*										
		(.17)										
Abstract-ness	Human	−.16*	.22*									
		(−.24)	(.29)									
Stability	Human	.11*	.59*	−.12*								
		(.14)	(.67)	(−.18)								
Importance	Human	.14*	.35*	−.11*	.47*							
		(.18)	(.41)	(−.17)	(.61)							
Base rate	Human	.34*	.46*	−.05	.33*	.36*						
		(.45)	(.54)	(−.08)	(.43)	(.47)						
Observa-bility	GPT-4o	**.67***	.08*	−.18*	.02	.07*	.23*					
		(.77)	(.08)	(−.24)	(.02)	(.08)	(.27)					
Social desirability	GPT-4o	.16*	**.95***	.18*	.63*	.37*	.40*	.11*				
		(.18)	**(.96)**	(.24)	(.71)	(.43)	(.47)	(.11)				
Abstract-ness	GPT-4o	−.19*	.14*	**.23***	.32*	.41*	−.01	−.28*	.20*			
		(−.22)	(.14)	**(.31)**	(.37)	(.48)	(.01)	(−.29)	(.20)			
Stability	GPT-4o	−.06	.22*	−.20*	**.59***	.41*	.06	.00	.28*	.38*		
		(−.07)	(.22)	(−.26)	**(.67)**	(.48)	(.07)	(.00)	(.28)	(.39)		
Importance	GPT-4o	−.08*	.15*	−.13*	.35*	**.58***	.11*	.02	.22*	.56*	.55*	
		(−.09)	(.15)	(−.17)	(.40)	**(.67)**	(.13)	(.02)	(.22)	(.57)	(.56)	
Base rate	GPT-4o	.22*	.70*	.04	.42*	.36*	**.66***	.15*	.67*	.01	.12*	.13*
		(.25)	(.71)	(.05)	(.47)	(.42)	**(.78)**	(.15)	(.67)	(.01)	(.12)	(.13)

Note. Zero-order correlations are Pearson product-moment correlations.*$p < .05$. Attenuation was calculated using ICC(2,k) as reliabilities. Correlations in parentheses are disattenuated for inter-rater unreliability.

GPT-4o replicated most of the correlation pattern among human ratings for social desirability, and also performed considerably well on four of the other properties (.71). The only exception was abstractness, for which GPT-4o's performance results were weaker (.23).

Given the high convergence with human ratings, our findings suggest that GPT-4o is capable of serving as a viable substitute for human raters when estimating social desirability, at least within conventional WEIRD (Heinrich et al., 2010) samples. Sentiment seems to be a central characteristic of communication (Dodds et al., 2015; Liu, 2012), and by extension, social desirability may be easier for LLMs to learn, potentially explaining the better results observed. However, GPT-4o rated more extremely than humans for social desirability, which could limit findings. Further, we would still advise caution when using synthetic ratings of the other five descriptor properties, as GPT-4o might occasionally conflate or misinterpret them (e.g., base rate and social desirability; see Table 6).

Limitations

Considerable effort went into reviewing participants' free responses to derive the final list of 876 descriptors. In hindsight, a few (e.g., "good-looking") might have been better excluded, though this likely had no impact given the large sample size. We advise readers intending to use the data set to carefully select descriptors that fit their research goals.

The reliability of the human ratings in the present study was lower than in some previous studies (e.g., Leising et al., 2014) for descriptor properties other than social desirability. On the one hand, this might be due to the online data collection, which could have produced lower data quality. On the other, we believe that the measures we took to control careless responding were effective and that the lower reliability could also have other, unproblematic causes (e.g., participant heterogeneity).

Although the study used descriptions of targets that differed in knowing and liking, potentially limiting neutral descriptors, this likely had minimal impact, as similar bimodal desirability patterns were found in studies using other methods for sampling descriptors (e.g., Anderson, 1968; Condon et al., 2022; Dumas et al., 2002). Further, the definition of several descriptor properties remains somewhat ambiguous. Systematic attempts at conceptual clarification would be needed. For example, abstractness and stability are both concerned with generality across situations and time, making it difficult to distinguish whether raters judge the breadth of behaviors implied by a descriptor or its temporal persistence. Future empirical studies should investigate potential effects of varying rating instructions on inter-rater reliability. Developing and then using consensual instructions for ratings of descriptor properties would enable clearer comparisons among studies.

As only German descriptors were used, replication in other languages is needed. Such studies would test generalizability and aim to produce reusable, rated descriptor sets in additional languages.

Outlook

The present study yielded a large and putatively representative sample of person descriptors, reliably rated for six key properties. The data is available for reuse in person perception research, enabling identical replication studies with several independent descriptor subsamples, and thus more generalizable conclusions. In addition, future research may use this data in the context of Multi-Level Profile Analysis (Biesanz, 2010; Wessels et al., 2020) for studying interactions between properties and perceiver-, target-, and dyadic effects.

There seems to be room for improvement in terms of how well LLM are able to capture the meaning of the different descriptor properties and distinguish them from one another. The present study only determined how well LLMs could emulate human judgments of descriptor properties, using the latter as the gold standard against which the former were validated. However, it would be quite possible that LLMs are actually able to capture descriptor properties *better* than a sample of human raters can. Investigating this would require comparative studies in which synthetic and human ratings compete with each other in predicting some criterion variable (e.g., model fit in predicting actual person judgments). LLMs bear promise in this regard because the effects of using different prompts to obtain ratings may be easily determined using them, but far less easy using samples of human raters.

Author Note

Not applicable.

Acknowledgements

We thank Johannes Zimmermann for feedback, Paula Knischweski for testing the reproducibility of the R-script, and Nele Freyer for reviewing the initial list of descriptors.

Author Contributions

Moritz Leistner: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

Björn E. Hommel: Formal analysis, Resources, Software, Writing – review & editing.

Leon P. Wendt: Formal analysis, Software, Validation, Writing – review & editing.

Daniel Leising: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Writing – review & editing.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: All funding was provided by internal resources and came from the chair of Assessment & Intervention of TU Dresden.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ORCID iDs

Moritz Leistner https://orcid.org/0009-0004-5737-1643
Björn E. Hommel https://orcid.org/0000-0002-7375-006X
Leon P. Wendt https://orcid.org/0000-0003-2229-2860
Daniel Leising https://orcid.org/0000-0001-8503-5840

Data Availability Statement

All data, preregistrations and materials can be found in the OSF folder.

Supplemental Material

Supplemental material for this article is available online. Depending on the article type, these usually include a Transparency Checklist, a Transparent Peer Review File, and optional materials from the authors.

Notes

Not applicable.

References

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

Arifin, W. N. (2023). Sample size calculator (web). Retrieved from. https://wnarifin.github.io/

Bäckström, M., Björklund, F., & Larsson, M. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality*, *43*(3), 335–344. https://doi.org/10.1016/j.jrp.2008.12.013

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. https://doi.org/10.48550/arXiv.2104.12250

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Biesanz, J. C. (2010). The social accuracy model of interpersonal perception: Assessing individual differences in perceptive and expressive accuracy. *Multivariate Behavioral Research*, *45*(5), 853–885. https://doi.org/10.1080/00273171.2010.519262

Bleidorn, W. (2012). Hitting the Road to Adulthood: Short-Term Personality Development During a Major Life Transition: Short-Term Personality Development During a Major Life Transition. *Personality and Social Psychology Bulletin*, *38*(12), 1594-1608. https://doi.org/10.1177/0146167212456707 (Original work published 2012)

Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine*, *21*(9), 1331–1335. https://doi.org/10.1002/sim.1108

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*, *55*(5), 1–18. https://doi.org/10.3758/s13428-022-01921-5

Burgoon, E. M., Henderson, M. D., & Markman, A. B. (2013). There are many ways to see the forest for the trees. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *8*(5), 501–520. https://doi.org/10.1177/1745691613497964

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). University of Toronto.

Condon, D. M., & Weston, S. J. (2022). Personality trait descriptors: 2,818 trait descriptive adjectives characterized by familiarity, frequency of use, and prior use in psycholexical research. *Journal of Open Psychology Data*, *10*(1), 1. https://doi.org/10.5334/jopd.57

Cottrell, C. A., Neuberg, S. L., & Li, N. P. (2007). What do people desire in others? A sociofunctional perspective on the importance of different valued characteristics. *Journal of Personality and Social Psychology*, *92*(2), 208–231. https://doi.org/10.1037/0022-3514.92.2.208

Della Rosa, P. A., Catricalà, E., Vigliocco, G., & Cappa, S. F. (2010). Beyond the abstract—concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. *Behavior Research Methods*, *42*(4), 1042–1048. https://doi.org/10.3758/brm.42.4.1042

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., Mitchell, L., Harris, K. D., Kloumann, I. M., Bagrow, J. P., Megerdoomian, K., McMahon, M. T., Tivnan, B. F., & Danforth, C. M. (2015). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(8), 2389–2394. https://doi.org/10.1073/pnas.1411678112

Duden. (2020). *Die deutsche Rechtschreibung. Das umfassende Standardwerk auf der Grundlage der aktuellen amtlichen Regeln*. ISBN 783411040186.

Dumas, J. E., Johnson, M. 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. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology*, *37*(2), 90–93. https://doi.org/10.1037/h0058073

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology*, *52*(2), 409–418. https://doi.org/10.1037/0022-3514.52.2.409

Hampson, S. E., John, O. P., & Goldberg, L. R. (1986). Category breadth and hierarchical structure in personality: Studies of asymmetries in judgments of trait implications. *Journal of Personality and Social Psychology*, *51*(1), 37–54. https://doi.org/10.1037/0022-3514.51.1.37unkel

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29. https://doi.org/10.1038/466029a

Hommel, B. E. (2023). Expanding the methodological toolbox: Machine-based item desirability ratings as an alternative to human-based ratings. *Personality and Individual Differences*, *213*(2023), 112307. https://doi.org/10.1016/j.paid.2023.112307

Hussain, Z., Binz, M., Mata, R., & Wulff, D. U. (2023). A tutorial on open-source large language models for behavioral science. https://osf.io/f7stn

John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancement and the role of narcissism. *Journal of Personality and Social Psychology*, *66*(1), 206–219. https://doi.org/10.1037/0022-3514.66.1.206

Leising, D., Borgstede, M.-, Burger, J., Zimmermann, J., Bäckström, M., Oltmanns, J., Freyer, N., Wiedenroth, A., Knischewksi, P., & Connelly, P. (2025). Why do judgments on different person-descriptive attributes correlate with one another? A conceptual analysis with relevance for Most psychometric research. *Collabra: Psychology*, *11*(1), 133683. https://doi.org/10.1525/collabra.133683

Leising, D., Ostrovski, O., & Borkenau, P. (2012). Vocabulary for describing disliked persons is more differentiated than vocabulary for describing liked persons. *Journal of Research in Personality*, *46*(4), 393–396. https://doi.org/10.1016/j.jrp.2012.03.006

Leising, D., Scharloth, J., Lohse, O., & Wood, D. (2014). What types of terms do people use when describing an individual's personality? *Psychological Science*, *25*(9), 1787–1794. https://doi.org/10.1177/0956797614541285

Leising, D., Scherbaum, S., Locke, K. D., & Zimmermann, J. (2015). A model of "substance" and "evaluation" in person judgments. *Journal of Research in Personality*, *57*(2015), 61–71. https://doi.org/10.1016/j.jrp.2015.04.002

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *112*(3), 474–490. https://doi.org/10.1037/pspp0000100

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

OpenAI. (2024). *GPT-4o (May 13 version) [large language model]*. https://platform.openai.com/docs/models/gpt-4o

Peabody, D. (1967). Trait inferences: Evaluative and descriptive aspects. *Journal of Personality and Social Psychology, 7*(4, Pt.2), 1–18. https://doi.org/10.1037/h0025230

Unkelbach, C., Fiedler, K., Bayer, M., Stegmüller, M., & Danner, D. (2008). Why positive information is processed faster: The density hypothesis. *Journal of Personality and Social Psychology*, *95*(1), 36–49. https://doi.org/10.1037/0022-3514.95.1.36

Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine*, *17*(1), 101–110. https://doi.org/10.1002/(sici)1097-0258(19980115)17:1<101::aid-sim727>3.0.co;2-e

Wessels, N. M., Zimmermann, J., Biesanz, J. C., & Leising, D. (2020). Differential associations of knowing and liking with accuracy and positivity bias in person perception. *Journal of Personality and Social Psychology*, *118*(1), 149–171. https://doi.org/10.1037/pspp0000218

Wessels, N. M., Zimmermann, J., Biesanz, J. C., & Leising, D. (2025). A nuanced perspective on how item features are associated with different forms of agreement. *Collabra: Psychology*, *11*(1), 127423. https://doi.org/10.1525/collabra.127423

Williams, J. E., Satterwhite, R. C., & Saiz, J. L. (1998). The importance of psychological traits: A cross-cultural study. https://ci.nii.ac.jp/ncid/BA4184489X

Wood, D. (2015). Testing the lexical hypothesis: Are socially important traits more densely reflected in the English lexicon? *Journal of Personality and Social Psychology*, *108*(2), 317–335. https://doi.org/10.1037/a0038343

Wood, D., & Wortman, J. (2011). Trait means and desirabilities as artifactual and real sources of differential stability of personality traits. *Journal of Personality*, *80*(3), 665–701. https://doi.org/10.1111/j.1467-6494.2011.00740.x

Wood, J. K., Anglim, J., & Horwood, S. (2021). 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/10.1177/08902070211012