MISS Measurement Instruments for the Social Sciences

PsychOpen GOLD

Advances in Methodology

Don't Pull Any Old Personality Taxonomy From the Shelf: The Performance of Historical and Sample Derived Taxonomies in Extracting Personality Information From Text

Johannes A. Karl [1,2], Ronald Fischer [3]

[1] *Department of Psychology, University of Zurich, Zurich, Switzerland.* **[2]** *School of Psychology, Victoria University of Wellington, Wellington, New Zealand.* **[3]** *Cognitive Neuroscience and Neuroinformatics Unit, Institute D'Or for Research and Education, São Paulo, Brazil.*

Measurement Instruments for the Social Sciences, 2026, Vol. 8, Article e16869, https://doi.org/10.5964/miss.16869

Received: 2025-02-02 • **Accepted:** 2026-01-15 • **Published (VoR):** 2026-04-16

Handling Editor: Marco Perugini, University of Milan Bicocca, Milan, Italy

Corresponding Author: Johannes A. Karl, Binzmuehlestrasse 14, 8050 Zurich, Switzerland. E-mail: Johannes.karl@psychologie.uzh.ch

Supplementary Materials: Data, Materials [see Index of Supplementary Materials]

Abstract

Substantial efforts have been made to develop comprehensive taxonomies of personality traits in many languages. Nevertheless, given that what is important and salient in individuals' lived experience might be changing over time, this raises the question about the long-term usefulness of 'off-the-shelf' taxonomies developed decades ago. In the current study we used a bottom-up approach to create a large sample-specific taxonomy of personality terms. We subsequently examined the overlap and sensitivity of this taxonomy compared to an established trait taxonomy in the same language. Overall, we found that the two taxonomies only showed limited overlap with a pronounced divergence in emotionality (Neuroticism) and social aspects (Agreeableness) of personality. In addition to this, we found that while the personality assessment extracted from self-descriptions using the established taxonomy showed alignment with participants' self-rated personality, especially Extraversion, Agreeableness, and Neuroticism, the sample-specific taxonomy showed a significantly greater alignment between text-based and self-rated personality. In summary, our current study highlights the need to extend our thinking about the psycholexical hypothesis, moving away from assumptions of time invariant language encoding to more explicitly recognizing temporal and sample-specific dynamics underpinning the expression and use of personality trait terms.

Keywords

lexical hypothesis, text-based personality assessment, text mining, sample specific taxonomy, Big Five

In 1884, Francis Galton (1949) famously asked: 'Can we discover landmarks of character to serve as bases for a survey, or is it altogether too indefinite and fluctuating to admit of measurement?' (pp. 179–180). Galton suggested that relevant moral faculties are 'so intermixed that they are never singly in action' (p. 181), yet he suggested it is possible to identify the most 'conspicuous aspects of the character'. To do so, Galton examined many pages of Roget's Thesaurus and estimated that it contained a fully one thousand words expressive of character. With this casual statement, he started an active field of inquiry known today as the psycholexical hypothesis (Ashton, Lee, Perugini, et al., 2004). There is a general consensus that a small number of factors can be used to describe human personality (Ashton & Lee, 2005; De Raad et al., 2010; Goldberg, 1993; Saucier et al., 2014). One of the core assumptions is that human communities will encode salient and important information about individual traits and character features in single terms in each language. Based on this assumption, taxonomies have been created that can be used to ask respondents to rate targets (Allport & Odbert, 1936; Ashton, Lee, & Goldberg, 2004; Goldberg, 1992; Norman, 1967; Saucier, 1994). However, language is dynamic and semantic content of words as well as the co-associations of individual words change over time (Xu et al., 2021). Consequently, it is important to ask whether taxonomies developed at some point in time with specific communities retain their usefulness over time. This is particularly important and interesting in the current social media environment with easier and wider access to user-created text that could be analysed with taxonomies as an unobtrusive measure of personality assessment (Boyd & Pennebaker, 2017; Suedfeld et al., 2011). Yet, this presumes a relative time-invariance of the taxonomies, an assumption which requires examination. We report the development of a theory-driven bottom-up English taxonomy in one specific sample of native English speakers and compare self-ratings based on this sample-specific taxonomy with both a commonly used off-the-shelf taxonomy and survey-based self-ratings.

Psychological Taxonomies to Capture Personality Traits

The first comprehensive taxonomy in English was developed by Allport and Odbert (1936). Norman (1967) empirically identified a five-factor structure based on a reduced list of taxonomy-derived ratings. Over time a consensus emerged that five or six factors are sufficient to describe the main variability underlying both self- and other ratings (Ashton, Lee, & Goldberg, 2004; Saucier & Goldberg, 1996). The most extensive of these taxonomies is the 1,710 personality-descriptive adjective list compiled by Goldberg (Goldberg, 1982), which has been a foundation for a number of factor-analytical studies (Ashton, Lee, & Goldberg, 2004). This list was given to undergraduate students in the US and Australia (total $N = 310$). Although initial analyses suggested up to seven factors,

the five and six factor solutions have been most widely used and the highest loading terms have been used as empirical markers for personality traits (Thalmayer et al., 2021; but see also Saucier & Iurino, 2020). Our work is guided by the five-factor solution differentiating Conscientiousness, Agreeableness, Neuroticism, Openness/Intellect and Extraversion (Goldberg, 1993; McCrae & John, 1992). We prefer this structure for our purposes because it is more parsimonious in describing the higher-order structure of broad personality domains and because the sixth factor (Honesty-Humility) tends to split off from within a broader Agreeableness factor (De Raad et al., 2010, 2014).

Language and Semantic Change

One interesting question is whether the factor structure of this taxonomy has remained stable over the last 40 years and across samples. Taking some hints from emotion research (Xu et al., 2021), emotion terms do change in their semantic meaning, as indicated by changing co-word associations in naturally occurring text over the last 100 years. Why should we be concerned with such changes? First, to the extent that personality traits have some biological foundation (DeYoung, 2014; McAdams & Pals, 2006), we should expect stability over time. At the same time, what and how we communicate important information is subject to cultural modification and transformation encoded in language (Bernardes et al., 2025; Christiansen & Chater, 2016). This argument is compatible with both cross-cultural and anthropological research suggesting that information is conveyed in locally relevant ways (and thereby temporally bound), which could result in changed factor structures. Such dynamics are relevant for any sample specific structures, which also applies to factor analysis models which reveal sample specific factor solutions. Therefore, the question of replicability of such structures across samples and time periods can provide important insights into the time and sample variant and invariant components of personality structure.

Second, with increasing availability of text via social media that could be used for personality assessment at a distance (Eichstaedt et al., 2021) and the generation of large language model and chatbots (Cutler & Condon, 2023; Fischer et al., 2023), one promising approach has been to rely on a bottom up analysis of text and then correlate any individual terms or combinations of terms with self-rated personality traits (Boyd & Pennebaker, 2017). For example, the open vocabulary approach has mapped word usage in Facebook status updates to personality self-ratings (Kern et al., 2014). This requires identification of relevant terms. Language as a communication tool is group and age specific, with slang and ideographic word use serving as identity badges to demark group membership along social and age specific boundaries (Nortier & Svendsen, 2015). As standard survey development exercises continue being informed by taxonomies within the lexical tradition (Thalmayer et al., 2021), it is important to study which personality terms are used by individuals from a specific sample to prevent incorrect results or conclusions.

Our interest is in identifying terms that our participants consensually use and understand to convey personality trait relevant information. We used definitions of the Big Five and asked participants to think of terms that they may use when describing an individual that is high or low on that particular trait. By using this approach, we use an explicit elicitation strategy which is transparent and participant driven and therefore, bottom up. Only terms that are salient for describing an individual with those theoretically meaningful characteristics are likely to be produced. Furthermore, by triangulating the word usage across our sample, we gain insights into the relative distribution of terms in this specific sample. Although the use of person-derived terms may seem anachronistic in times of Large Language Models and machine-learning approaches to natural text for extracting possible personality markers (Giannini et al., 2024), we believe that the participant-driven approach is a distinct strength over these computational methods. Generally, machine learning and transformer-based approaches need to be trained on specific corpora and rely on a number of unexamined assumptions about the stability and representativeness of the training text (Bender et al., 2021; Hu et al., 2025; Mehrabi et al., 2021), turning them into virtual 'black-boxes' that reduce transparency and replicability. For example, to what extent are descriptions of venues good proxies of personality descriptions, unless we want to make certain assumptions about how humans describe both other humans and venues (see Giannini et al., 2024)? The proliferation of training data derived from popular models such as ChatGPT also raises the risk of deterioration of signal (Shumailov et al., 2024). Using human-derived data with explicit instructions and verifying the consensus and overlap between participants provides a transparent option for creating a list of terms that participants use to describe each other. Furthermore, the black-box nature of transformed based models makes it difficult to study semantic drift over time given that it is often not easily understood and comparable how scores are calculated. Therefore, once replicated across samples and across time periods, our method offers a distinct advantage for more fine-grained contextual analyses.

To evaluate how relevant those terms are, we utilized an open writing task in which participants had to describe themselves. This task allows us to compare the performance of our sample-specific taxonomy with the published taxonomy. We extracted terms from these self-descriptions and mapped them to a) our bottom-up theory-driven taxonomy and b) the taxonomy by Ashton and colleagues. We also compared the relative correlation of these two text-based scores with each other and with self-ratings using a standard psychology questionnaire (Soto & John, 2017). Considering possible temporal change, we also examined overlap in these taxonomies—what terms are used by our sample when describing individuals high or low on a personality trait and how well are they captured by classic taxonomies developed roughly 40 years ago.

Method

Participants

Our sample consisted of 317 participants who took part exchange for course credit (mean age = 19.22 years, *SD* = 3.08; 77.9% female). The sample size was determined by logistical constraints of running the study within the context of a university degree. Our actual sample size allowed for a minimum detectable correlation (80% power, $\alpha = .05$) of $r = 0.14$. Our study was open to self-enrolment by the target population until a pre-specified cut-off date. All de-identifiable data is available on the open science framework (https://osf.io/hn69f/overview). The personal narratives of subjects are removed due to ethical considerations of anonymity.

Measures

BFI-2

We used the BFI-2 to assess personality (Soto & John, 2017). The overall scale had 60 items and participants reported their agreement with each item on a 1-(Disagree strongly) to 5-(Agree strongly) Likert-scale. Example items were "I am someone who is outgoing, sociable" and "I am someone who is compassionate, has a soft heart". All dimensions showed high reliability: $\omega_{Extraversion}$: .849[.826, .872], $\omega_{Agreeableness}$: .828[.802, .854], $\omega_{Conscientiousness}$: .850[.828, .873], $\omega_{Neuroticism}$: .909[.895, .922], $\omega_{Openness}$: .817[.790, .845].

Self-Description

Participants were prompted with the following statement for a self-description: "We would like to ask you to describe yourself in 500 words (this is roughly a single page or 2000 characters). Who are you as a person?" The average response was 1853.09 (SD = 182.83) characters long (min = 1301, max = 2000). This self-description task was presented in a counterbalanced fashion with the BFI-2 across participants.

Personally – Relevant Personality Terms

To create a sample level taxonomy, participants were lastly prompted for each of the five factors of personality to submit 10 terms (5 positive and 5 negative) which they would use to label a person either high or low on this trait. These trait descriptions were based on definitions and descriptions of the big five in the literature (Bernardes et al., 2022; DeYoung, 2014; Fischer, 2017; Soto & John, 2017). For example, for extraversion participants were prompted: "Persons with high scores on Extraversion tend to be sociable and energetic in social interactions, they get a lot of energy out of being with others. What words would you use to describe such individuals to your friends?". This task was always presented last. Overall, participants provided 3900 unique personality terms. We

excluded terms with less than two characters or a frequency below three. This filtering resulted in a list of 703 unique terms. As participants were able to nominate a term for multiple categories or different participants naming a term for different categories, we assigned personality terms to a category based on their most frequent mention. We dropped terms with equal nominations across dimensions. Our final taxonomy of terms consisted of 671 terms that were commonly mentioned and clearly attributable to one of the five factors of personality. We show the full taxonomy in the supplementary material in STable 1. We show the terms excluded due to non-distinguishable double-nominations in STable 2. Terms were equally distributed across positive (N = 328) and negative terms (N = 354). Examining distributions across positive and negative terms, participants provided significantly more negative Agreeableness and negative Openness terms ($\chi^2(4) = 13.21$, $p < .010$; see Table 1).

Existing Personality Taxonomy

We used the 1710 taxonomy (Ashton, Lee, & Goldberg, 2004) as a starting point, but we only used trait terms that were unambiguously loading with loadings > .30 and cross loadings < .20 in the original study. This resulted in 405 terms. Exploratory analyses with larger word sets (which included more cross-loading terms and lower loading terms) did not substantively change the performance of this taxonomy (see footnote 2). In the final version used here, these terms were equally distributed across positive (N = 198) and negative terms (N = 207). Positive and negative terms were equally distributed within traits ($\chi^2(4) = 3.496$, p = .479; see Table 1). Importantly, this taxonomy had substantially less Openness and Neuroticism terms (see Table 1) compared to the other traits.

Table 1

Terms in Each Taxonomy by Positive and Negative Direction

Direction	A	C	E	N	O
Sample Derived					
Negative	92	66	53	60	83
Positive	58	75	59	76	60
Historical 1710					
Negative	57	53	51	35	11
Positive	48	63	52	24	11

Extraction of Term-Based Personality From Text

To extract the personality data from text, we first created a list based on each term corpus for the two taxonomies using the *quanteda* package. Prior to extraction we removed punctuation, numbers, symbols, common English stopwords, and coerced all words to

lowercase to allow for unambiguous matching. For each participant we extracted the total number of words used and the personality terms matched in each taxonomy. To increase the comparability across participants we normalized each personality score for each participant by dividing it by the number of total words written and centred the score around their mean usage of personality terms.

Results

Overlap of Taxonomy Terms

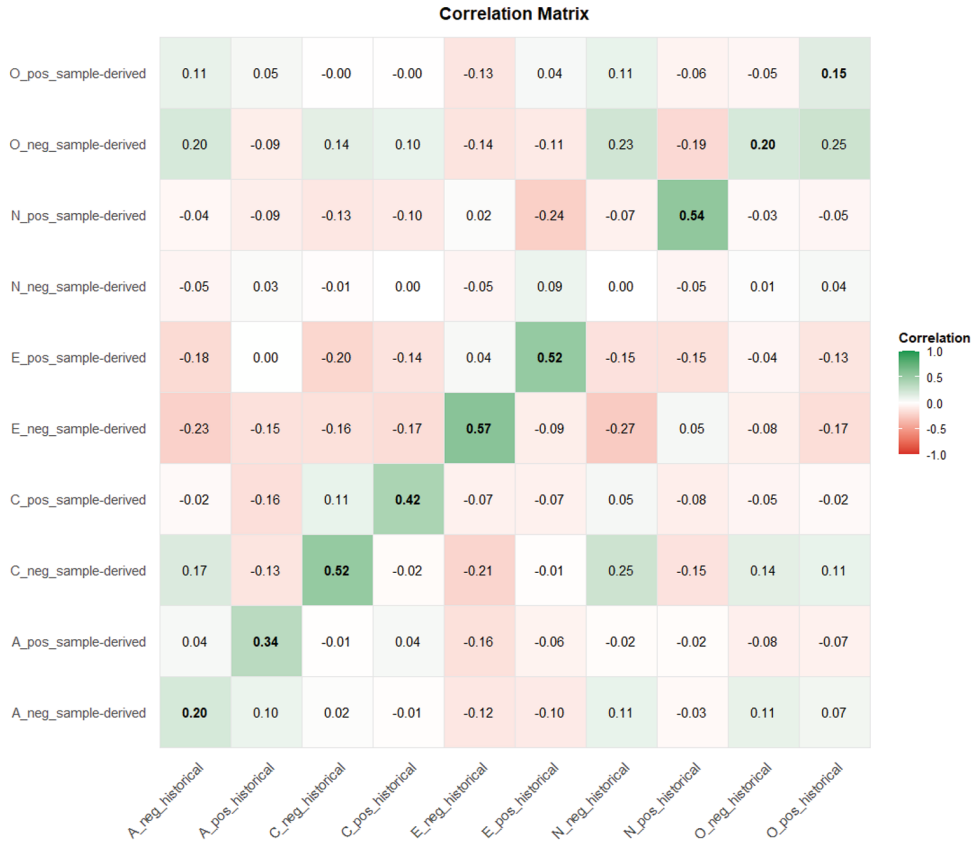
We first examined the shared terms between our sample specific taxonomy and the off-the-shelf taxonomy. Overall, we found that the taxonomies had an overlap of 19.75%. The taxonomies had the greatest overlap for Openness (27.27%), Conscientiousness (23.28%), and Extraversion (21.36%), but we found a lower overlap for Neuroticism (15.25%) and Agreeableness (15.24%). We show the overlapping terms in Supplementary Table 3 (see Karl & Fischer, 2026).

Overlap of Extracted Personality Between Taxonomies

To examine the overlap in extracted personality between the taxonomies we correlated the score of each participant across dimensions and term directions between the taxonomies. On average the taxonomies correlated at $r = .28$ and scores were significantly positively correlated across the taxonomies except for negative Neuroticism (we show all correlations in Figure 1, correlation tables are available on the OSF). While some dimensions such as Extraversion had a substantial correlation $r > .50$ for both positive and negatively valenced terms, others such as openness had a smaller correlation. For Neuroticism, positively valenced terms correlated quite strongly, whereas negatively valenced terms showed virtually no correlation. Taken together these patterns imply that the extracted personality differed substantially across the taxonomies which might be due to the terms not shared between the taxonomies. Similar taxonomy-based effects have been reported previously (Bernardes et al., 2025; Fischer et al., 2020). In other words, the terms included in taxonomies are idiosyncratic and specific taxonomy usage may result in different patterns for the same data set.

Figure 1

Correlation Between Sample-Derived Scores and Historical 1710 Taxonomy Scores



Self-Report — Text-Based Personality Congruence

To examine the similarity of self-ratings and text-based personality assessment we examine the correlation between participants scored personality according to each taxonomy within the text and their self-rating on the BFI-2. For ease of interpretation this was split by positive and negative terms. To confirm the robustness of the difference in correlations for dependent groups we used Hittner et al.'s (2003) procedure. Hittner et al.'s (2003) modified Z-test is a statistical procedure designed to test whether two correlation coefficients derived from the same sample differ significantly from one another. This test is necessary when comparing dependent correlations because the correlations share a common variable, violating the independence assumption required for standard correlation comparison tests. The procedure accounts for the intercorrelation between

the variables being compared, adjusting the standard error to reflect the dependency structure in the data. This approach provides a more accurate assessment of whether observed differences in correlation magnitudes are statistically meaningful rather than due to chance.

As can be seen in Table 2, we found that while the off-the-shelf taxonomy showed small to medium correlations with self-rated personality ($Mean_{positive\ terms}$: .124, range: .04 to .21; $Mean_{negative\ terms}$:-.122, range: -.31 to .02), the sample specific taxonomy qualitatively outperformed it using positive and negative terms ($Mean_{positive\ terms}$: .194, range: .11 to .24, $Mean_{negative\ terms}$: -.118, range: -.27 to -.02). The correlations between sample-specific taxonomy scores and self-ratings significantly differed from the correlation between off-the-shelf taxonomy scores and self-ratings for positive C terms and positive O terms (all $p < .05$).

Table 2

Correlation of BFI Self-Ratings With Sample Derived or Historical Positive and Negative Terms in the Text-Based Extraction

Trait (self-report)	Sample-Derived Positive	Sample-Derived Negative	Historical 1710 Positive	Historical 1710 Negative	Sample-Derived Positive (Reduced)	Sample-Derived Negative (Reduced)
E	0.24***	-0.27***	0.21***	-0.31***	0.28***	-0.27***
A	0.11*	-0.12*	0.14**	-0.18***	0.11*	-0.13**
C	**0.18*** **	-0.10*	0.04	-0.09	0.17**	-0.10*
N	0.20***	-0.08	0.17**	-0.05	0.17**	-0.06
O	**0.24*** **	-0.02	0.06	0.02	0.16** [a]	-0.03

Note. Correlations in bold significantly differ at $p < .05$ between the sample derived and historical taxonomies. Columns marked 'reduced' exclude terms that can be found in the researcher provided trait descriptions.
[a] indicates significant differences from the original term-self report correlation.
***$p < .001$. **$p < .01$. *$p < .05$.

A final analysis was to compare the overall pattern of the correlations across positive and negative terms for each taxonomy with the self-report scores. The overall correlation of the pattern was $r = .91$. This suggests that the correlation pattern of taxonomies with self-ratings was highly similar, pointing towards problems with specific traits instead of overall non-comparability. In this regard, it was interesting to note that positive terms showed a greater tendency to pick up participants' self-rated personality.[1] Only E showed medium sized correlations for both positive and negatively valenced terms for both

1) We explored the difference between the full 1710 historical dictionary and our cleaned version. Overall, using the positive terms of the full version the 1710 historical dictionary showed a lower correlation with self-rated personality for Extraversion ($r = .11$, $p < .05$), Agreeableness ($r = .10$, $p = .07$), Conscientiousness ($r = .04$, $p = .39$) and Openness ($r = .01$, $p = .79$), but a higher relationship for Neuroticism ($r = .19$, $p < .001$) compared to the cleaned version.

taxonomies with self-reports. In contrast, N and O showed essentially zero correlations for the negative pole.

Robustness Checks

Half of the participants saw the Big Five measure before the free self-description task, which may have influenced their responses to either measure or subsequently nominated terms. To address the potential impact, we conducted five separate analyses. First, we examined the frequency responses between the sets of responses based on shared terms between the sets of participants that completed the free-text before and after the Big Five measures. We computed a rank-order correlation and found a significant correlation of .96, $p < .001$, indicating a high degree of similarity in the frequency of terms nominated which were shared. Second, when using all terms nominated and computing the similarity across the imbalanced set, we find a Jaccard similarity of 58.56% indicating a substantial overlap in the specific terms nominated (even when allowing for rare terms).

To examine if in the subsequent trait nomination participants only replicated terms from their self-description or the Big Five measure we ran two additional analyses. First, prior to examining that participant's self-provided personality terms were not overly overlapping with their self-descriptions, we extracted the ratio of terms nominated by participants that could also be found in their self-description. On average the overlap between self-provided terms and terms used in their description was 2.66 terms whereas the overlap with all terms provided by participants was 13.28 terms on average. This indicates that participants were substantially more likely to use terms in their self-description that were not found in their nominated terms later. Finally, to examine the possibility that participants were primed by our trait description to use specific terms, we examined the incidence ratio of a term being present in the researcher provided descriptor on its nomination by participants. Overall, we found that terms in the descriptor were nominated more often, with an incidence rate ratio of 4.19 (95% CI [3.98, 4.42], $p < .001$), suggesting that terms from existing personality descriptor lists were approximately four times more frequently nominated by participants compared to novel terms. Nevertheless, it is difficult to conclude if this is due to the prototypicality of the selected terms or general priming. Therefore, to examine the robustness of our analysis to the exclusion of the terms in our trait description, we reran the analysis excluding terms that could be found in the description of the trait (Table 2). We only found one significant difference with the relationship of personality ratings based on extracted terms with BFI self-report weaker for positive openness, but the correlation was still substantially higher compared to using the 1,710 terms.

Discussion

One of our major questions motivating the current research was whether sample specific taxonomies of personality are better at capturing participants' personality compared with self-reports than established off-the-shelf taxonomies. Overall, our study shows that sample specific taxonomies out-perform off-the-shelf taxonomies in capturing participants' personality, especially for Conscientiousness and Openness. This is not to say that off-the-shelf taxonomies do not present a valuable research tool, especially if no sample-based taxonomy can be created due to logistical reasons (e.g., all members of the study population are deceased).

Our results nevertheless point to a number of challenges in this broad area going forward. The correlation between personality scores extracted from text using previously published taxonomies and sample-specific taxonomies was relatively weak on average (r = .28), corresponding to a moderate effect size for individual difference research (Gignac & Szodorai, 2016). This may be somewhat disappointing but probably not surprising considering that the overlap in terms across taxonomies was less than 20% across all traits. Furthermore, the correlations between self-ratings using standard survey inventories and text-based scores were again low, with a slight advantage for sample-specific taxonomies. These patterns raise questions on a) whether self-reports using surveys or text-based scores provide complementary or distinct information, b) which language terms within text may be most indicative of personality traits, c) whether some traits are better detectable via text and others via self- (or other) reports and d) more broader concerns about determining the ground-truth in relation to human personality (Boyd et al., 2020; Boyd & Pennebaker, 2017). We will selectively discuss some of these issues next.

Concerning specific patterns that stood out and may speak to the four questions just mentioned: negative Openness and Neuroticism descriptors showed very low correlations with self-reports. These low correlations are contrasted with the comparatively high correlations for negative Extraversion. This pattern raises a few intriguing possibilities. Firstly, in lexical approaches terms are used as equally weighted in their indication of the construct, which contrasts with findings that people are more likely to use positive terms compared to negative terms. At the same time, rarer terms convey more information (Garcia et al., 2012). There is the possibility that this frequency - information density ratio of positive and negative terms varies across traits affecting the signal ratio. Alternatively, some researchers have highlighted the complex conceptual nature of Openness (Schwaba & Thalmayer, 2025) and the variability the trait behaviour link of Openness and Neuroticism (Soto, 2021), which might especially manifest in negations increasing the difficulty of signal detection. Another important point to consider is the number of terms available within a taxonomy, which may increase the ability to detect signals. For example, our sample-specific taxonomy contained more terms for Openness, which may have increased the ability to detect weak signals in text and this in turn increased the correlation with self-reports. Yet, removing marker terms significantly decreased

this correlation. This again points to the importance for future research to examine the information contained in marker terms vis-à-vis the breadth of personality traits.

Further, our results indicate that while samples agree on a substantial corpus of personality terms, a considerable portion of taxonomy entries may be idiosyncratic. Our sample was culturally similar to the samples which were used to derive the off-the-shelf taxonomy, yet our samples were separated by roughly 40 years. Some traits such as Neuroticism and Agreeableness showed a markedly larger shift in content and performance. To speculate why these traits might have shifted more, both are related to emotional content which might show an increased rate of change over time (Xu et al., 2021). Alternatively, socio-cultural changes might have resulted in a different conceptual construction of these terms. Especially in light of recent studies which show an accelerating rise of cognitive distortions which are related to both interpersonal and emotion-regulation (Bollen et al., 2021), we may expect larger divergences in socially and emotionally focused traits. This highlights the possibility that the seemingly greater change in Neuroticism and Agreeableness terms might be temporally specific and the emergence of different cultural patterns might dampen or exacerbate this trend.

In our current study we focused on the five-factor model of personality, yet this leaves open the question how other potential traits, such as Honesty-Humility within the HEXACO (Ashton, Lee, Perugini, et al., 2004) might perform. Honesty-Humility has been viewed as part of Agreeableness and has shown substantial correlations in some studies (De Raad et al., 2010). An interesting potential example of the ambiguity of meaning can be found in the way participants have labelled the term *honest* in our data, which has been equally classified as positive Agreeableness, negative Agreeableness, or negative Openness. In the full 1,710 wordlist the original sample rated this term equally as an indicator of Agreeableness and Conscientiousness.

Importantly, recent studies have challenged the universality of the big five theory (Fischer, 2017, 2021; Laajaj et al., 2019), suggesting that different trait structures may emerge in different social and ecological settings. Our approach suggests that the terms included within the taxonomies (or items within surveys) may not be representative of the traits within those specific samples. This issue has been identified as the problem of indicator relevance and representativeness (Fischer & Karl, 2019; Fontaine, 2005). The issue with the traditional lexical hypothesis is that it assumes time invariant information mapping. However, linguistic shifts do occur, and taxonomies are unlikely to remain stable. Examining the indicator relevance and representativeness problem from a lexical hypothesis perspective, we could argue that the lexical basis of this hypothesis is more aligned with temporally and sample-specific dynamic indicator-to- construct mappings. Moving away from assumptions of time invariant language encodings may open ways for a better understanding of what information is relevant to be passed on within specific language communities and how this information maps onto cognitive schema that people hold about socially relevant constructs. We believe that such an explicit recognition

of temporal and sample-specific information value can open important new insights into both personality structure and personality dynamics over time (Fischer & Rudnev, 2025).

Limitations and Future Research Directions

To allow for a comparison with established trait taxonomies, our current study was limited to one specific sample in one anglophone context which is culturally and linguistically similar to the original samples used to develop and validate the off-the-shelf taxonomy. This limits our insight on change and similarity in taxonomy performance to the English language. It would be important for future studies to extend this line of research using some of the recently developed trait term taxonomies in diverse language groups and study their performance with new samples within each language group. Similarly, our study represents a specific sample, which skews mostly female and is all university students. University students represent a relatively homogeneous group in terms of cognitive ability and socioeconomic status, which could influence both personality manifestations and self-descriptions. At one level, our approach therefore highlights the benefit of tailoring terms to a specific sample, but by necessity also limits the generalizability of our findings and the resultant taxonomies to other samples. We believe that this limitation highlights a major point of the current study, namely that while existing taxonomies of personality can pick up signals of personality from text, researchers working with specific samples might benefit from expanding these taxonomies by generating bottom-up trait descriptors to capture a clearer signal in their respective sample.

Our study is also limited by its cross-sectional nature and relying on a single sample. Although, we can get some insight into the change of personality descriptors in presumed culturally comparable cohorts over time, it would nevertheless be an important future avenue to examine the change of personality descriptors within and across samples.

Our study provides initial evidence that semantic drift may influence the performance of established trait taxonomies comparing a contemporary snapshot to an existing historical dictionary derived forty years ago with the aim to provide initial insight into potential drift. To extend our understanding of semantic change comparable data should be systematically collected on bottom-up personality descriptors across multiple time periods and cohorts to examine and validate the construction of personality categories over time. Such a temporally distributed datasets would allow the use of diachronic word embeddings (Hamilton et al., 2018; Kutuzov et al., 2018) to enable researchers to track systematic shifts in the meaning, usage, and semantic associations of trait terms across historical periods in a bottom up fashion, rather than imposing ahistorical trait definitions. By applying these computational approaches to personality-relevant vocabulary collected across multiple time points, future research could directly quantify the magnitude and nature of semantic drift in trait terms, determine which descriptors

remain stable versus which undergo substantial meaning shifts, and identify the cultural and linguistic factors that drive such changes.

Starting with a contemporary personality model such as the Big Five, we presuppose that this structure is applicable and relevant to our sample. Given the current evidence, it seems reasonable to assume the applicability in Western and highly educated samples (Laajaj et al., 2019; Soto & John, 2017; Thalmayer et al., 2022). At the same time, individuals in other cultural and linguistic contexts may share implicit personality structures that diverge from this Big Five model identified in student samples like ours, with either fewer or more factors (Cheung et al., 2001; Fischer, 2017; Gurven et al., 2013; Nel et al., 2012; Thalmayer et al., 2021). This clearly requires substantive additional work, in order to identify locally meaningful personality models as well as their relevant marker terms.

By conducting bottom-up analyses with human populations or by using computational methods to identify period-specific word embeddings, it may be possible to identify both more time invariant (e.g., models and markers that are relative insensitive to temporal changes) and time variant personality models and descriptors. Moving beyond human derived trait lists, researchers may start with seed words from person descriptions in text from in different temporal periods and compute word embeddings of key terms identified. These word embeddings can then be further queried to map systematic changes in valence, salience or breadth (see Baes et al., 2024). This approach aligns with an emerging historical psychological movement (Jackson & Atari, 2025) that seeks to understand how psychological constructs themselves evolve across historical periods, recognizing that personality traits are culturally and temporally situated phenomena (Du et al., 2024; Fischer et al., 2020). Critically, this historical approach enables more accurate and comprehensive study of psychological concepts by allowing researchers to examine temporal change and potentially time-invariant features together within a unified framework. Rather than treating historical variation as noise to be controlled away, this method treats both changing and stable aspects of personality as substantive phenomena worthy of investigation, thereby providing a more complete picture of how personality operates across both time and culture. To the extent that it is possible to identify systematic factors that influence the emergence and structuring of personality terms across time, this would open new opportunities for testing evolutionary models of personality. We see our study as a first stepping stone in this direction, which will require systematic replications and extensions across different cultural samples and time periods.

A further limitation that is shared by most lexical studies is the so-called ground-truth problem, that is, what scores can be considered to capture personality dynamics with the greatest accuracy and validity. We used self-report ratings as comparison standards, but other behaviour-based options need to be explored in future research (Boyd & Pennebaker, 2017). Finally, we focused on the five factor model, which leaves an open

question about stability and change in personality descriptors related to culture specific social-relational traits (Fetvadjiev et al., 2015).

Conclusion

In summary, our study shows that both off-the-shelf and sample-specific taxonomies can be used to extract personality information from narratives and self-descriptions, but a sample-specific taxonomy might be preferable as it exhibits greater sensitivity and shows more similar patterns to self-report measures. Our study demonstrates the need to move beyond the idea of one personality taxonomy per sample, but rather focus more study on how personality expression changes within samples over time to separate potential time-invariant descriptors of personality from descriptors idiosyncratic to a specific temporal instance of a sample.

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

Data Availability: For this article, data is freely available (see Karl & Fischer, 2022).

Supplementary Materials

For this article, the following Supplementary Materials are available:

- Data (see Karl & Fischer, 2022)
- Additional analyses and results. STable 1 presents the full participant-provided personality trait dictionary, listing terms generated by participants for each Big Five dimension (Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness) organised by positive and negative valence. STable 2 reports terms that were excluded from the dictionary due to double nominations across multiple facets. STable 3 lists overlapping terms between the participant-provided dictionary and existing personality taxonomies. The document also includes robustness checks examining task order effects on the convergent validity correlations, including a random-effects mini-meta-analysis and an accompanying forest plot (SFigure 1) (see Karl & Fischer, 2026)

Index of Supplementary Materials

Karl, J. A., & Fischer, R. (2022). *The performance of off-the-shelf and population derived lexica in extracting implicit personality from self-descriptions* [Data]. OSF. https://osf.io/hn69f

Karl, J. A., & Fischer, R. (2026). *Supplementary materials to "Don't pull any old personality taxonomy from the shelf: The performance of historical and sample derived taxonomies in extracting*

personality information from text" [Tables, figures]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.21826

References

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs, 47*(1), Article i–171. https://doi.org/10.1037/h0093360

Ashton, M. C., & Lee, K. (2005). The lexical approach to the study of personality structure: Toward the identification of cross-culturally replicable dimensions of personality variation. *Journal of Personality Disorders, 19*(3), 303–308. https://doi.org/10.1521/pedi.2005.19.3.303

Ashton, M. C., Lee, K., & Goldberg, L. R. (2004). A hierarchical analysis of 1,710 English personality-descriptive adjectives. *Journal of Personality and Social Psychology, 87*(5), 707–721. https://doi.org/10.1037/0022-3514.87.5.707

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., de Vries, R. E., Di Blas, L., Boies, K., & De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology, 86*(2), 356–366. https://doi.org/10.1037/0022-3514.86.2.356

Baes, N., Haslam, N., & Vylomova, E. (2024). *A multidimensional framework for evaluating lexical semantic change with social science applications*. arXiv. https://doi.org/10.18653/v1/2024.acl-long.76

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bernardes, G., Bozza, B., Motta, M., Mattos, P., & Fischer, R. (2025). Semantic meaning means a lot: Exploring the role of semantics in the development of a Big Five taxonomy. *Journal of Research in Personality, 115*, Article 104570. https://doi.org/10.1016/j.jrp.2024.104570

Bernardes, G., Fischer, R., & Motta, M. (2022). *Personality encoded in language: A theory-based dictionary in Brazilian Portuguese*. https://doi.org/10.17605/OSF.IO/QD4ET

Bollen, J., ten Thij, M., Breithaupt, F., Barron, A. T. J., Rutter, L. A., Lorenzo-Luaces, L., & Scheffer, M. (2021). Historical language records reveal a surge of cognitive distortions in recent decades. *Proceedings of the National Academy of Sciences of the United States of America, 118*(30), Article e2102061118. https://doi.org/10.1073/pnas.2102061118

Boyd, R. L., Pasca, P., & Lanning, K. (2020). The personality panorama: Conceptualizing personality through big behavioural data. *European Journal of Personality, 34*(5), 599–612. https://doi.org/10.1002/per.2254

Boyd, R. L., & Pennebaker, J. W. (2017). Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences, 18*, 63–68. https://doi.org/10.1016/j.cobeha.2017.07.017

Cheung, F., Leung, K., Zhang, J.-X., Sun, H.-F., Gan, Y.-Q., Song, W.-Z., & Xie, D. (2001). Indigenous Chinese personality constructs: Is the five-factor model complete? *Journal of Cross-Cultural Psychology, 32*(4), 407–433. https://doi.org/10.1177/0022022101032004003

Christiansen, M. H., & Chater, N. (2016). *Creating language: Integrating evolution, acquisition, and processing*. MIT Press.

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. https://doi.org/10.1037/pspp0000443

De Raad, B., Barelds, D. P. H., Levert, E., Ostendorf, F., Mlacić, B., Di Blas, L., Hrebícková, M., Szirmák, Z., Szarota, P., Perugini, M., Church, A. T., & Katigbak, M. S. (2010). Only three factors of personality description are fully replicable across languages: A comparison of 14 trait taxonomies. *Journal of Personality and Social Psychology, 98*(1), 160–173. https://doi.org/10.1037/a0017184

De Raad, B., Barelds, D. P. H., Timmerman, M. E., De Roover, K., Mlačić, B., & Church, A. T. (2014). Towards A Pan–Cultural Personality Structure: Input from 11 Psycholexical Studies. *European Journal of Personality, 28*(5), 497–510. https://doi.org/10.1002/per.1953

DeYoung, C. G. (2014). A cybernetic Big Five theory for personality psychology. *Personality and Individual Differences, 60*, S18. https://doi.org/10.1016/j.paid.2013.07.381

Du, A. H., Karl, J. A., Fetvadjiev, V., Luczak-Roesch, M., Pirngruber, R., & Fischer, R. (2024). Tracing the evolution of personality cognition in early human civilisations: A computational analysis of the Gilgamesh epic. *European Journal of Personality, 38*(2), 274–290. https://doi.org/10.1177/08902070231161869

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., Hagan, C. A., Tobolsky, V. A., Smith, L. K., Buffone, A., Iwry, J., Seligman, M. E. P., & Ungar, L. H. (2021). Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods, 26*(4), 398–427. https://doi.org/10.1037/met0000349

Fetvadjiev, V. H., Meiring, D., van de Vijver, F. J. R., Nel, J. A., & Hill, C. (2015). The South African Personality Inventory (SAPI): A culture-informed instrument for the country's main ethnocultural groups. *Psychological Assessment, 27*(3), 827–837. https://doi.org/10.1037/pas0000078

Fischer, R. (2017). *Personality, values, culture: An evolutionary approach*. Cambridge University Press. https://doi.org/10.1017/9781316091944

Fischer, R. (2021). Alternative four-factor structure of the Mini-IPIP in Thailand. *International Journal of Personality Psychology, 7*, 35–42. https://doi.org/10.21827/ijpp.7.37978

Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology, 10*, Article 1507. https://doi.org/10.3389/fpsyg.2019.01507

Fischer, R., Karl, J. A., Luczak–Roesch, M., Fetvadjiev, V. H., & Grener, A. (2020). Tracing personality structure in narratives: A computational bottom–up approach to unpack writers,

characters, and personality in historical context. *European Journal of Personality, 34*(5), 917–943. https://doi.org/10.1002/per.2270

Fischer, R., Luczak-Roesch, M., & Karl, J. A. (2023). *What does ChatGPT return about human values? Exploring value bias in ChatGPT using a descriptive value theory.* arXiv. https://doi.org/10.48550/arXiv.2304.03612

Fischer, R., & Rudnev, M. (2025). From MIsgivings to MIse-en-scène: The role of invariance in personality science. *European Journal of Personality, 39*(4), 662–673. https://doi.org/10.1177/08902070241283081

Fontaine, J. R. J. (2005). Equivalence. In *Encyclopedia of Social Measurement* (Vol. 1, pp. 803–813).

Galton, F. (1949). The measurement of character. In W. Dennis (Ed.), *Readings in general psychology* (pp. 435–444). Prentice-Hall. https://doi.org/10.1037/11352-058

Garcia, D., Garas, A., & Schweitzer, F. (2012). Positive words carry less information than negative words. *EPJ Data Science, 1*(1), Article 3. https://doi.org/10.1140/epjds3

Giannini, F., Marelli, M., Stella, F., Monzani, D., & Pancani, L. (2024). Surfing the OCEAN: The machine learning psycholexical approach 2.0 to detect personality traits in texts. *Journal of Personality, 92*(6), 1602–1615. https://doi.org/10.1111/jopy.12915

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Goldberg, L. R. (1982). From Ace to Zombie: Some explorations in the language of personality. *Advances in Personality Assessment, 1*, 203–234.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *The American Psychologist, 48*(1), 26–34. https://doi.org/10.1037/0003-066X.48.1.26

Gurven, M., von Rueden, C., Massenkoff, M., Kaplan, H., & Lero Vie, M. (2013). How universal is the Big Five? Testing the five-factor model of personality variation among forager-farmers in the Bolivian Amazon. *Journal of Personality and Social Psychology, 104*(2), 354–370. https://doi.org/10.1037/a0030841

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). *Diachronic word embeddings reveal statistical laws of semantic change.* arXiv. https://doi.org/10.48550/arXiv.1605.09096

Hittner, J. B., May, K., & Silver, N. C. (2003). A Monte Carlo evaluation of tests for comparing dependent correlations. *The Journal of General Psychology, 130*(2), 149–168. https://doi.org/10.1080/00221300309601282

Hu, T., Kyrychenko, Y., Rathje, S., Collier, N., van der Linden, S., & Roozenbeek, J. (2025). Generative language models exhibit social identity biases. *Nature Computational Science, 5*(1), 65–75. https://doi.org/10.1038/s43588-024-00741-1

Jackson, J. C., & Atari, M. (2025). Historical psychology: How the events of yesterday shaped the minds of today. *Current Research in Ecological and Social Psychology, 9*, Article 100247. https://doi.org/10.1016/j.cresp.2025.100247

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Dziurzynski, L., Ungar, L. H., Stillwell, D. J., Kosinski, M., Ramones, S. M., & Seligman, M. E. P. (2014). The online social self: An open vocabulary approach to personality. *Assessment, 21*(2), 158–169. https://doi.org/10.1177/1073191113514104

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1384–1397). Association for Computational Linguistics. https://aclanthology.org/C18-1117/

Laajaj, R., Macours, K., Pinzon Hernandez, D. A., Arias, O., Gosling, S. D., Potter, J., Rubio-Codina, M., & Vakis, R. (2019). Challenges to capture the big five personality traits in non-WEIRD populations. *Science Advances, 5*(7), Article eaaw5226. https://doi.org/10.1126/sciadv.aaw5226

McAdams, D. P., & Pals, J. L. (2006). A new Big Five: Fundamental principles for an integrative science of personality. *The American Psychologist, 61*(3), 204–217. https://doi.org/10.1037/0003-066X.61.3.204

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys, 54*(6), Article 115. https://doi.org/10.1145/3457607

Nel, J. A., Valchev, V. H., Rothmann, S., van de Vijver, F. J. R., Meiring, D., & de Bruin, G. P. (2012). Exploring the personality structure in the 11 languages of South Africa. *Journal of Personality, 80*(4), 915–948. https://doi.org/10.1111/j.1467-6494.2011.00751.x

Norman, W. T. (1967). *2800 personality trait descriptors: Normative operating characteristics for a university population*. University of Michigan, Dept. of Psychology.

Nortier, J., & Svendsen, B. A. (Eds.). (2015). *Language, youth and identity in the 21st century: Linguistic practices across urban spaces*. Cambridge University Press. https://doi.org/10.1017/CBO9781139061896

Saucier, G. (1994). Mini-Markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment, 63*(3), 506–516. https://doi.org/10.1207/s15327752jpa6303_8

Saucier, G., & Goldberg, L. R. (1996). The language of personality: Lexical perspectives on the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 21–50). Guilford Press.

Saucier, G., & Iurino, K. (2020). High-dimensionality personality structure in the natural language: Further analyses of classic sets of English-language trait-adjectives. *Journal of Personality and Social Psychology, 119*(5), 1188–1219. https://doi.org/10.1037/pspp0000273

Saucier, G., Thalmayer, A. G., Payne, D. L., Carlson, R., Sanogo, L., Ole-Kotikash, L., Church, A. T., Katigbak, M. S., Somer, O., Szarota, P., Szirmák, Z., & Zhou, X. (2014). A basic bivariate structure of personality attributes evident across nine languages. *Journal of Personality, 82*(1), 1–14. https://doi.org/10.1111/jopy.12028

Schwaba, T., & Thalmayer, A. G. (2025). Openness/Intellect: A unique trait requires unique considerations. *Personality and Social Psychology Review*. Advance online publication. https://doi.org/10.1177/10888683251377227

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature, 631*(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y

Soto, C. J. (2021). Do links between personality and life outcomes generalize? Testing the robustness of trait–outcome associations across gender, age, ethnicity, and analytic approaches. *Social Psychological & Personality Science, 12*(1), 118–130. https://doi.org/10.1177/1948550619900572

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Suedfeld, P., Cross, R. W., & Brcic, J. (2011). Two years of ups and downs: Barack Obama's patterns of integrative complexity, motive imagery, and values. *Political Psychology, 32*(6), 1007–1033. https://doi.org/10.1111/j.1467-9221.2011.00850.x

Thalmayer, A. G., Job, S., Shino, E. N., Robinson, S. L., & Saucier, G. (2021). ǂŪsigu: A mixed-method lexical study of character description in Khoekhoegowab. *Journal of Personality and Social Psychology, 121*(6), 1258–1283. https://doi.org/10.1037/pspp0000372

Thalmayer, A. G., Saucier, G., & Rotzinger, J. S. (2022). Absolutism, relativism, and universalism in personality traits across cultures: The case of the Big Five. *Journal of Cross-Cultural Psychology, 53*(7–8), 935–956. https://doi.org/10.1177/00220221221111813

Xu, A., Stellar, J. E., & Xu, Y. (2021). Evolution of emotion semantics. *Cognition, 217*, Article 104875. https://doi.org/10.1016/j.cognition.2021.104875