Stage 1 Registered Report (not peer reviewed, only pilot data reported)

Language models accurately infer correlations between psychological items and scales from text alone

Björn E. Hommel[1, 2*] and Ruben C. Arslan[1]

[1] Wilhelm Wundt Institute of Psychology, Leipzig University, Germany

[2] magnolia psychometrics GmbH, Germany

Björn E. Hommel https://orcid.org/0000-0002-7375-006X

Ruben C. Arslan https://orcid.org/0000-0002-6670-5658

* Corresponding author: Björn E. Hommel (bjoern.hommel@uni-leipzig.de)

Björn E. Hommel and Ruben C. Arslan contributed equally to this paper.

Online supplement:

- Statistical reports and interactive plots: https://rubenarslan.github.io/surveybot3000/
- OSF (Code & Data): https://osf.io/z47qs/,
- App: https://huggingface.co/spaces/magnolia-psychometrics/synthetic-correlations

Abstract:

Many behavioural scientists do not agree on core constructs and how they should be measured. Different literatures measure related constructs, but the connections are not always obvious to readers and meta-analysts. Many measures in behavioural science are based on agreement with survey items. Because these items are sentences, computerised language models can make connections between disparate measures and constructs and help researchers regain an overview over the rapidly growing, fragmented literature. Our fine-tuned language model, the SurveyBot3000, accurately predicts the correlations between survey items, the reliability of aggregated measurement scales, and intercorrelations between scales from item positions in semantic vector space. In our pilot study, the out-of-sample accuracy for item correlations was .71, .86 for reliabilities, and .89 for scale correlations. in a preregistered study, we will investigate whether the performance of our model generalises to measures across behavioural science.

Introduction

Behavioural science struggles to be cumulative in part because scientists in many fields fail to agree on core constructs (Bainbridge et al., 2022; Sharp et al., 2023). The literature silos, which consequently develop, can appear unconnected but pursue the same phenomena under different labels (see e.g., grit and conscientiousness; Credé et al., 2017).

One reason why connections are lacking is the asymmetry inherent in measure and construct validation: adding novel constructs to the pile is easier than sorting through it. Investigators can easily invent a new ad-hoc measure and benefit reputationally if a new construct becomes associated with their name (Elson et al., 2023; Flake & Fried, 2020). By contrast, finding out whether a purported new construct or measure is redundant with the thousands of existing ones is cumbersome and can cause conflict with other researchers (Bainbridge et al., 2022; Elson et al., 2023). The same holds for replicating construct validation studies and reporting evidence of overfitting or other problems (Hussey et al., 2024; Kopalle & Lehmann, 1997).

Untangling the "nomological net"—a term coined by Cronbach and Meehl (1955) to describe the relationships between measures and constructs—has become increasingly difficult given the growing number of published measures (Anvari et al., 2024; Elson et al., 2023). Conventional construct validation methods, though effective in mapping these relationships, do not scale to, for instance, the thousands of measures that might be related to neuroticism. To tackle this problem, Condon and Revelle (2015; see also Condon, 2017; Condon et al., 2017) have championed the Synthetic Aperture Personality Assessment in which survey participants respond to a small random selection of a large set of items from the personality literature. Over time, as the sample size grows, this procedure allows estimating pairwise correlations between all items. Although the approach is efficient, each new item requires thousands of participants to answer the survey before it can be correlated with all existing items. Hence, the approach cannot be used to quickly evaluate new proposed scales. What is missing is an efficient way to prioritise, to prune the growth in constructs and measures and to sort through the disorganised pile of existing measures.

Natural language processing could provide this efficiency. In the social and behavioural sciences, subjective self-reports are one of the predominant forms of measurement. The textual nature of survey items lends itself to natural language processing. Recently, transformer models have become the state-of-the-art in language models (Vaswani et al., 2017), displaying proficiency in understanding and generating text. They have dramatically reduced the costs of many tasks and chores, notably in programming and generating images from verbal prompts. Although capabilities for natural language generation are currently more visible in the public eye through the use of chat-like interfaces, they are backed by capabilities in natural language understanding (e.g., classifying or extracting features from text).

On a technical level, this understanding is implemented by the so-called encoder block, which processes input text and encodes it as a high-dimensional numeric vector. The vector representation of a word like “party” in the resulting semantic vector space is context-dependent.

The same word will yield a different vector representation if it occurs in the statement "I am the life of the party" compared to "I always vote for the same party". The encoder block's ability to contextualise words is crucial for recognizing the nuances of language. At heart, the efficiency of the transformer model can largely be attributed to its self-attention mechanism (Vaswani et al., 2017). As the name suggests, it is loosely analogous to the executive function in human cognition. Instead of "memorising" an entire corpus of text, word by word, the attention mechanism weights the relevance of words in a context window for a given target word.

Transformer models excel in transfer learning, that is, they adapt to new tasks easily (Tunstall et al., 2022). Following the pre-training stage, which establishes a base level of linguistic expertise, the models can undergo domain adaptation, which involves training the model on a corpus of text specifically curated for the task at hand. In a process called fine-tuning, the model then retains its originally learned weights but learns to carry out a specific task, such as text classification. Essentially, the model builds on the fundamental knowledge acquired during pre-training to adapt to specialised tasks, even with limited training data. This concept is known as few-shot learning. High-quality annotated training data is key for the domain adaptation that turns generalists into specialists.

Using linguistic information to scaffold scientific models has a long history in personality psychology, where the lexical hypothesis states that more important personality characteristics are more likely to be encoded as words. To find important personality dimensions, researchers had human subjects rate themselves on prominent adjectives, or *items*, identified systematic correlations between items, and applied factor analytic techniques to reduce the number of dimensions. The most popular organising framework, the Big Five, was distilled from personality-descriptive items in this manner (Digman, 1990).

Pre-transformer era attempts to use semantic features of items to predict associations between measurement scales using latent semantic analysis have demonstrated moderate utility (Arnulf et al., 2014; Larsen & Bong, 2016; Rosenbusch et al., 2020; Hernandez & Nie, 2023). As the ability of computerised language models to capture meaning has grown, researchers have sought to directly quantify relationships between adjectives from textual data (Cutler & Condon, 2022), to assign items to constructs (Fyffe et al., 2024; Guenole et al., 2024), to directly predict item responses (Abdurahman et al., 2024; Argyle et al., 2023) and quantify open-ended answers to questions (Kjell et al., 2019, 2024).

Wulff & Mata (2023) used large language models (LLMs) to map survey items to vector space and predict empirical item correlations. They tested various transformer models for their ability to predict properties of psychological inventories. They observed a correlation of $r = .22$ between the semantic similarities of items as judged by OpenAI's ada-002 model (Greene et al., 2022) and the item correlations estimated in empirical data, with accuracy improving when aggregating vectors to the scale level. Their work shows large language models can approximately infer item correlations and outperform latent semantic analysis. However, their approach relied on pre-

trained models that were not adapted to the domain of survey items and do not appreciate that empirical item correlations are often negative because of negation. This approach cannot be expected to unlock the latent ability of the models, but rather to give a lower bound of their usefulness. At the same time, pre-trained models can overfit to their training data. Because OpenAI's large language models obtain knowledge from scraping large quantities of internet text, they presumably have seen items from existing measures co-occur in online studies and public item repositories. The results for survey items that inadvertently were part of the training data can lead to more optimistic results than could be expected for novel items.

We have adapted a sentence transformer model to the domain of survey response patterns and trained our model, the SurveyBot3000, to place items in vector space. The distances between item pairs in vector space produce what we will call *synthetic* item correlations, scale correlations, and reliabilities. These synthetic estimates can potentially help to cheaply evaluate measures and constructs. We plan to validate that the SurveyBot3000 can approximately infer empirical item correlations in data not used to train the model. To do so, we will preregister the model's synthetic estimates before collecting empirical data using a sample of survey participants. Based on our pilot study, we predict that the model will exhibit substantial accuracy in inferring empirical item correlations (r = .71, 95% CI [.70;.72]), and even higher accuracy in inferring latent correlations between scales (r = .89 [.88;.90]) and in inferring reliability coefficients (r = .86 [.76;.95]). We detail our predictions in our Design Table.

If our validation confirms that synthetic estimates are accurate, our model can be put to work in multiple areas. Synthetic correlations will always require careful follow-up with empirical data, but they can be used to search and prioritise. Authors can use our model as a semantic search engine to find existing constructs and measures and avoid reinventions. Synthetic correlations could be used as inputs for more realistic *a priori* power analyses. Scientific reviewers can use it to flag optimistic reliability coefficients and unstable factor structures, especially when researchers have not validated an ad-hoc measure out-of-sample yet. Generally, discrepancies between reported estimates and LLM-based synthetic estimates can motivate greater attention to replication and construct validation. Finally, meta-scientists and measurement researchers can use the model to start sorting through the pile of tens of thousands existing constructs and measures (Anvari et al., 2024; Elson et al., 2023).

As a showcase, we have made the model available as an app on Huggingface. Researchers can enter item texts and the app will generate synthetic item correlations, scale correlations and reliability coefficients. The app contains a prominent cautionary note to discourage researchers from taking the synthetic estimates at face value before further validation has occurred.

Methods

Materials, data, and code for the present study are available through the Open Science Framework: https://osf.io/z47qs/. Data pre-processing, model training, and statistical analyses were conducted using Python (version 3.10.12; Van Rossum & Drake, 2009), R (version 4.3.1; R Core Team, 2023), with an Nvidia GeForce RTX 2070 Super GPU, using the CUDA 11.7.1 toolkit (NVIDIA et al., 2022).

Ethics information

The planned research complies with the ethics guidelines by the German Society for Psychology (Berufsverband Deutscher Psychologinnen und Psychologen, 2022). Data used in model training were collected by third parties, as shown in the online supplemental section (https://osf.io/z47qs/). Participants in the validation study will be recruited from the crowdsourcing platform *Prolific*, and compensated at a median wage of $12 per hour. Informed consent will be obtained from all human respondents. Ethics approval for the validation study has been requested from the Institutional Review Board (IRB) at Leipzig University and will be amended once the design is finalised after review. All necessary support is in place for the proposed research.

Pre-trained language model

Our preliminary work has focused on improving the predictions of item correlations with sentence transformer models using high-quality training corpora for domain adaptation. We modified a LLM to generate synthetic item correlations by fine-tuning a pre-trained sentence transformer model (Reimers & Gurevych, 2019). Unlike conventional transformer models used in natural language understanding tasks which produce vector representations of individual tokens (i.e., basic linguistic units, such as words or syllables), sentence transformers produce vector representations for longer sequences of text (e.g., sentences).

Sentence transformers—specifically the bi-encoder architecture used throughout this research—work by using two parallel LLMs that process text inputs independently but share the same structure and parameters. The central idea behind these models is to capture the semantic essence of a sentence. One method to accomplish this is by pooling (e.g., averaging) the contextualised token vectors for each of the two models and then combining them. The underlying neural network then learns these combined representations by predicting sentence similarities, for instance using natural language inference data. In natural language inference, a given text (i.e., the premise) is evaluated based on its relation to another text (i.e., the hypothesis), classified as either contradicting, entailing, or being neutral to it. The network's output layer consists of three neurons, each representing one of these classes. The model's learning effectiveness is assessed using cross-entropy loss, with improvements in sentence vector representation achieved through backpropagation. Interested readers are referred to Reimers & Gurevych (2019), as well as Schroff et al. (2015) for further details on bi-encoders. Accessible in-

depth introductions to transformer models and deep neural networks can be found in Hussain et al. (2023) and Hommel et al. (2022).

We chose the *all-mpnet-base-v2* model (hereafter referred to as the "SBERT model" for further fine-tuning from the Hugging Face model hub (*Hugging Face Model Hub*, n.d.), based on its commendable performance across 14 benchmark datasets (*Pretrained Models — Sentence-Transformers Documentation*, n.d.). This pre-trained model is a sentence-transformer adaptation of the *mpnet-base* model (Song et al., 2020), initially trained on 160 gigabytes of English language text, including Wikipedia, BooksCorpus, OpenWebText, CC-News, and Stories. The SBERT model places sentences in a 768-dimensional semantic vector space. Distances in this Euclidean space can be computed using, for instance, cosine similarity. In our case, we hypothesised that the cosine similarity between the vector representations of any two survey items (e.g., personality statements) should correspond to the correlation coefficients obtained from survey data.

Domain adaptation and fine-tuning

We fine-tuned the pre-trained model in two stages. In the first stage, we trained the model to distinguish between semantically opposing concepts. In the second stage, we trained the model to predict pairwise item correlations, using survey data. Figure 1 depicts the multi-staged training procedure.

Stage 1: Polarity calibration Although cosine similarity spans from -1 to 1, negative coefficients are rarely produced when comparing vector representations of sentences (cf. the croissant shape of the top left plot in Figure 2). This limitation primarily arises because the high-dimensional vector representation of sentences encodes a range of abstract linguistic features, many of which tend to be positively correlated across text sequences. This poses a challenge in accurately predicting correlations for items of opposing scale polarities, such as those on the introversion-extraversion continuum. To illustrate, when assessing cosine similarity between items from the pre-trained model, the item "I am the life of the party" produces comparable coefficients with "I make friends easily" (Θ = .32) and "I keep in the background" (Θ = .35). This occurs even though the last item represents the polar opposite of the first item.

We fine-tuned the pre-trained model with the goal of maximising the cosine distance between vector representations of opposing concepts. We achieved this by augmenting the Stanford Natural Language Inference corpus (SNLI version 1.0, see also Supplementary Note 3; (Williams et al., 2018) for our purposes. SNLI comprises around 570,000 sentence pairs, each labelled for textual entailment as either contradiction, neutral, or entailment. We re-labelled each sentence pair by additionally assigning a magnitude to the semantic relationship. We let the pre-trained SBERT model generate the cosine similarity of the sentence pair (e.g., "the moon is shining" and "it is a sunny day", Θ = .46), but assigned a negative direction if the pair was labelled as contradictory (e.g., Θ = -.46). Hence, our new criterion combined the magnitude and direction of the similarity, capturing various forms of negation in the process. The fine-tuned model was

then trained to predict this new criterion, so that it would learn that similar sentences have negative cosine similarities when one sentence negates or contradicts the other (see Supplementary Note 6 for more detailed evaluation metrics).

Stage 2: Domain adaptation We found that the SBERT model's predictions of item correlations were skewed by the presence of non-trait-related text in the item stems. Specifically, we identified a tendency for item correlations to be overestimated in statements containing the same adverbs of frequency. For example, the phrase "I *often* feel blue" from the depression facet of the NEO-PI-R in the IPIP exhibits similar cosine similarity to the two items "I feel that my life lacks direction" (Θ = .28) and "I *often* forget to put things back in their proper place" (Θ = .26), even though the first item is also from the depression facet while the second is from the orderliness facet.

To address this, we aimed to fine-tune the model to focus on text segments that convey information relevant to psychological traits and their similarity. This adjustment aimed to enhance the model's accuracy in identifying and processing trait-relevant language and to teach it about personality structure, thus improving the validity of its synthetic correlations. We compiled training data from 29 publicly available online repositories (see Supplementary Note 4). Our inclusion criteria for the corpus mandated that raw item-level data be available, a minimum sample size of $N \geq 300$, the use of a rating scale as response format, and clear mapping of item stems to variable names in the datasets. In pre-processing, we retained pairwise Pearson coefficients from the lower triangular matrix across all datasets and cleaned and standardised item stems. Further details on the preprocessing of data can be found on the OSF (https://osf.io/bfhzy). For cross-validation purposes, we distributed each item pair among training, validation, and test partitions, adhering to an 80-10-10 split. To avoid overfitting, we ensured that all items were unique to their partition. This led to the exclusion of a substantial portion of our training data. Specifically, from the initial pool of 204,424 item pairs, we retained 90,424 pairs. Of these, we randomly allocated 74,339 pairs (82%) to the training partition, 6,832 pairs (8%) to the validation partition, and 9,253 pairs (10%) to the test partition. To mitigate the risk of the model learning idiosyncratic characteristics inherent to the dataset —item stems within a dataset are more likely to exhibit resemblance than between datasets— we used an additional holdout dataset. This dataset comprised 87,153 item pairs obtained from Bainbridge et al. (2022) thereby providing a robust measure for evaluating the model's generalizability to novel English language items about personality and related individual differences. To ensure the integrity of the holdout dataset, any items not exclusive to it were eliminated from the training, validation, and test partitions.

We optimised the hyperparameters for fine-tuning the model using the Optuna library in Python (version 3.1.1; Akiba et al., 2019), with a focus on enhancing the model's ability in predicting item correlations within the test partition. Details of the final hyperparameter selection are available in the online supplemental material (https://osf.io/b5ua7).

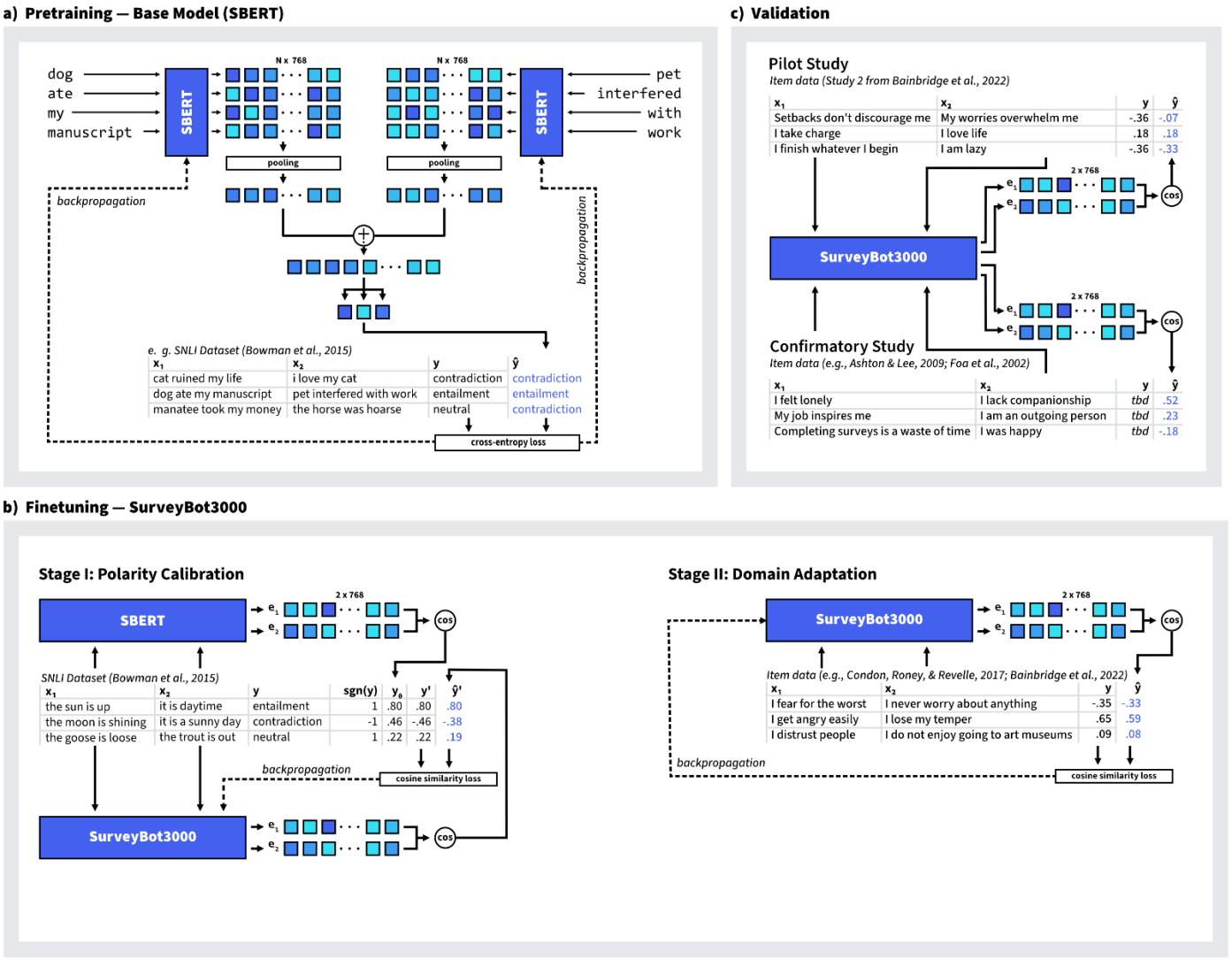

Figure 1. Multi-staged training procedure for the SurveyBot3000, which produces synthetic estimates of inter-item correlations.

Pilot study

We found that the SurveyBot3000 model was highly accurate for all partitions of the *curated* corpus. Empirical inter-item correlations and synthetic correlations were accurately predicted in the test set $r = .69$ ($df = 9{,}251$; 95% CI [.67, .70]) and in the validation set $r = .71$ ($df = 6{,}830$; 95% CI [.70, .72]). That accuracy was high in both test and validation set shows the model's strong generalizability within the corpus.

The SurveyBot3000 model was then tested using 87,153 item pairs obtained from Bainbridge et al. (2022), the holdout dataset we withheld from the training process to avoid over-fitting. Adjusted for sampling error in the empirical data (see Supplementary Note 1), the model's synthetic correlations predicted the empirical inter-item correlations with an accuracy of r = .71 (95% CI [.70;.72], manifest correlation r = .67, Figure 2). This consistency with the test-set performance shows the model's ability to generalise beyond the idiosyncratic properties of the data seen in training. Figure 2 shows the prediction of item correlations through semantic similarity, as estimated by the SBERT and SurveyBot3000 models. The SBERT model had substantially lower accuracy in predicting inter-item correlations in our holdout (manifest $r = .19$ [.18;.19]).

We further investigated the model's ability to predict scale reliabilities, which can be calculated from inter-item correlation matrices. Given that scales are typically designed to exhibit high internal consistency, we observed limited variability in the internal consistency measures across the 107 scales and subscales in the holdout dataset. Empirical Cronbach's alpha values had a mean of .75 ($SD = .10$) and ranged from .35 to .93. When new scales are designed, reliability varies more widely. We therefore circumvented the problem of restricted variance by randomly sampling items to create 200 additional, varied scales. We omitted random scales whose empirical Cronbach's alpha estimate was negative. We found that synthetic reliability estimates were highly accurate at $r(253) = .86$, 95% CI [.74, .94] (manifest $r = .82$ [.78;.85]. Again, the SBERT model had substantially lower accuracy (manifest $r = .07$ [-0.04;.18]). Accuracy was lower when we excluded the randomly formed scales (manifest $r = .63$ [.50;.73]), as expected owing to the restricted range in the real scales (SD = .10 compared to SD = .23 in the combined set).

We subsequently investigated the model's validity for scale-level predictions using the holdout dataset. We averaged the vector representations of all items in each scale and then computed the cosine similarity of these averaged vectors. The convergence between empirical and synthetic scale correlations was remarkably high, exhibiting an accuracy of $r(6{,}245) = .89$ [.88, .90] (manifest correlation r = .87 [.86;.87]). In other words, our fine-tuned LLM explained 80% of the latent variance in scale intercorrelations, based on nothing but semantic information contained in the items. Again, the SBERT model had substantially lower accuracy (manifest $r = .33$ [.30;.35]).

In summary, the LLM-based synthetic estimates closely approximated the empirical inter-item and inter-scale correlations as well as reliability estimates and were robust to the checks detailed in Supplementary Note 2. Comparing predictions between the datasets used in this pilot study leads us to expect that the effects are robust and will generalise to new, previously unseen English-language items.

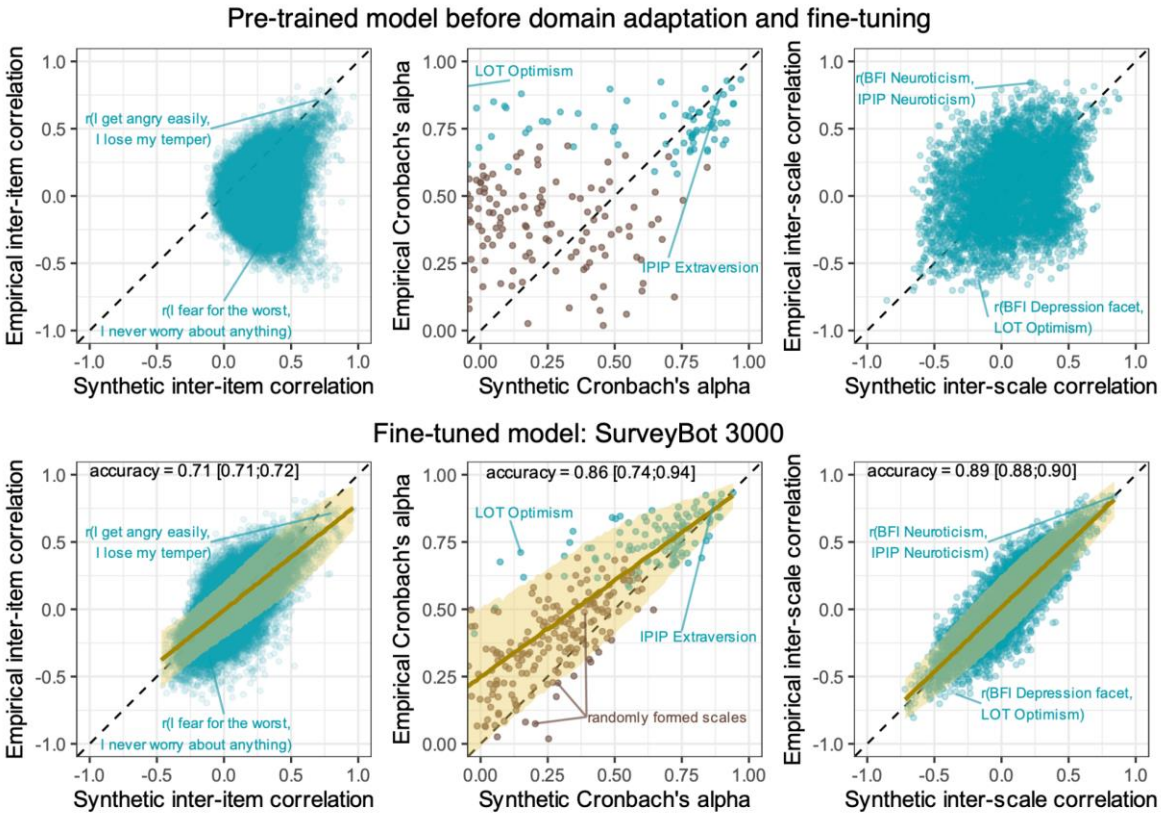

Figure 2. Scatter plots of the synthetic and empirical estimates. We show N=87,153 item pair correlations, N=255 scale reliabilities, and N=6,245 scale pair correlations for the pre-trained SBERT model (first row) and the fine-tuned SurveyBot3000 model (second row). The yellow line and shaded yellow region show the prediction and the 95% prediction interval for the latent outcome according to a Bayesian multi-membership regression model that allowed for heteroskedasticity and sampling error. Because the empirical estimates are estimated with sampling error, which the model adjusts for, fewer than 95% of dots are in the shaded prediction interval. Brown dots in the middle column show randomly combined scales, which we used to increase variance in the criterion. For reliabilities, three real and 19 randomly combined scales with negative synthetic alphas are not shown for ease of presentation.

Design

The primary objective of our research is to test the generalisability of our model in predicting human response patterns within survey data, that is, empirical item and scale correlations, as well as scale reliabilities. Our model's initial training data and our holdout represent a limited subset of the broader universe of survey items, with a skew towards personality psychology. We designed our validation study to challenge the model's capabilities by sampling from a more varied array of psychological measures. We plan to collect empirical data from a large online sample of English-speaking US Americans, similar to most of the studies in our training data. Participants will fill out the scales in random order, with item order randomised in each scale. While we anticipate a modest reduction in effect size during Stage 2 compared to the outcomes observed in the pilot study, we expect that the LLM-based synthetic estimates will still be sufficiently accurate to be useful. We present a Design Table summarising our methods and benchmarks.

Measures

To identify appropriate measures for our study, we conducted a comprehensive search of the APA PsycTests database. Our inclusion criteria for selecting scales were: a) utilisation of rating scales as the response format, b) items composed in the English language, c) scales developed within the last 30 years to minimise confounding factors related to changes in the English language, d) measures applicable to the general population, thus excluding scales only applicable to narrow target demographics such as adoptive parents or particular professional groups, e) measures applicable to a broad domain, thus excluding scales designed to rate specific consumer products or specific social attitudes, and f) freely accessible, non-proprietary measures. These criteria were mainly intended to make it feasible to have an unselected sample respond to most items. Within these constraints, we sampled scales to cover a wide range of measures used in the social and behavioural sciences.

We did not always use all items in a scale, so that we would be able to have participants respond to a large number in a scale. We included measures from industrial/organisational psychology, such as the Utrecht Work Engagement scale, measures from social psychology such as the Moral Foundations Questionnaire, from developmental psychology, such as the Revised Adult Attachment Scale, from clinical psychology, such as the Center for Epidemiological Studies Depression Scale, from emotion psychology, such as the positive and negative affect schedule, from personality psychology, such as Honesty-Humility in the HEXACO-60, and from other social sciences, such as the Attitudes Toward AI in Defence Scale and the Survey Attitude Scale. A full list of all scales can be found in Supplementary Note 5 and all items were deposited on OSF. In all, we plan to have participants answer 246 items distributed across 81 scales and subscales.

For all measures, we adapted the response format to a 6-point Likert scale from *strongly disagree* to *strongly agree*. We decided that a more uniform presentation was more important

than a perfectly faithful rendering of the original scale. In addition, our current model is unaware of differing response formats and cannot account for them.

Sampling Plan

We used simulations to determine our number of scales, items, and survey participants. We want to precisely estimate the accuracy with which our synthetic estimates can approximate empirical estimates of inter-item and inter-scale correlations. Sampling error at the participant level affects the standard error with which we estimate empirical inter-item and inter-scale correlations and therefore would bias our accuracy estimates downward. To estimate empirical individual item correlations, we use an online panel provider to collect a representative US sample of $N = 450$, before exclusions. We will limit participant recruitment to participants who have an approval rate exceeding 99% and have participated in at least 20 previous studies according to the sample provider, *Prolific*. We will pay participants regardless of whether they fail attention checks or complete the survey too quickly. In our planned analyses, we will then estimate the accuracy of our manifest synthetic estimates for latent, error-free empirical estimates (see Supplementary Note 1).

From the APA PsycTests corpus, we plan to sample 246 items, which can be aggregated to 57 scales consisting of at least three items. We assumed we would retain a sample of at least $n = 400$ after exclusions. With the resulting 30,135 unique item pairs, we should be able to infer the accuracy of our synthetic inter-item correlations to a precision (standard error) of ±0.004, according to our simulations. If we supplement our 57 scales with 200 randomly constituted scales, we should be able to infer the accuracy of our synthetic reliability estimates to a precision of ±0.03. With the resulting 1,558 unique scale pairs, without scale-subscale pairs, we should be able to infer the accuracy of our synthetic inter-scale correlations to a precision of ±0.007. The achieved precision is sufficient to detect even subtle deterioration in accuracy compared to our pilot study estimates.

Analysis Plan

We will follow Goldammer et al.(2020) and Yentes (2020) recommendations for identifying and excluding participants exhibiting problematic response patterns (e.g., careless responding). Accordingly, participants will be excluded if any of the following thresholds are exceeded: a) longstring (≥ .40 SD above mean), b) multivariate outlier statistic using Mahalanobis distance (≥ .50 SD above mean), c) psychometric synonyms ($r < .60$), d) psychometric antonyms ($r \geq -.40$), e) even-odd-index (≥ .20 SD above mean).Then, we will compute all empirical inter-item correlations, inter-scale correlations, and reliabilities. Inter-item correlations will be Pearson's product-moment correlations. We aggregate scales as the means of their items after reversing reverse-coded items. Inter-scale correlations are then computed as manifest Pearson's product-moment correlations. Reliability will be estimated with the Cronbach's alpha coefficient based on

inter-item correlation. We have uploaded synthetic estimates of the SBERT model and the SurveyBot3000 model for all of these coefficients to the OSF. The code for our preregistered analyses will mirror the code from our pilot study, including the robustness checks detailed in Supplementary Note 2. We will freeze both code and point predictions as part of our preregistration. After data collection, we will merge empirical and synthetic estimates. We will then compute accuracies, that is the correlations between synthetic and empirical estimates, disattenuated for the standard error of the empirical estimates using a Bayesian errors-in-variables model, which allows for heteroskedastic accuracy (see Supplementary Note 1). We will also report the prediction error for all three quantities, as well as a plot similar to Figure 2. We will also report manifest accuracies and the accuracy of the SBERT model, which we will use as a benchmark (see Design Table).

Table 1. Design Table

Question	Hypothesis	Sampling plan	Analysis Plan	Interpretation given to different outcomes
How accurate are LLM-based synthetic inter-item correlations?	The synthetic estimates will exhibit an accuracy of $r = .71$ for the empirical inter-item correlation coefficients obtained from survey data, as estimated in our Bayesian multi-membership regression model.	246 items. With the resulting 30,135 unique item pairs, we should be able to estimate accuracy with a precision of ±0.004. A representative sample of N=400 will be drawn to estimate empirical correlations.	A correlation between synthetic and empirical estimates, disattenuated for the sampling error in the empirical estimates.	If the accuracy matches (i.e. *±.02)* that found in our pilot study, this is evidence that the model generalises well to novel survey items, including those outside personality psychology. In the unlikely case that the accuracy exceeds that found in our pilot study, we would carefully discuss why, including the potential that crowdworkers use LLMs to respond. If the accuracy deteriorates to within 60% of the r in the pilot, the model may still be useful but should be applied with caution when item content is unlike the training data. We will examine and discuss performance across subfields to understand the deterioration. Retraining the model on a broader corpus would be indicated for future research. If the accuracy deteriorates to below 60% of the r in the pilot, our model does not generalise well. Retraining with a broader corpus would be needed before recommending the model for wider use. If the accuracy of our model is reduced below the accuracy of the pre-trained model, our model training procedure overfit despite our precautions. The model should not be recommended for practical use
How accurate are LLM-based synthetic reliability coefficients (for scales consisting of at least three items)?	The synthetic estimates will exhibit an accuracy of $r = .86$ for the empirical Cronbach's alpha coefficients obtained from survey data, as estimated in our Bayesian regression model.	As above. With the available 57 scales, supplemented by 200 randomly formed scales, we should be able to estimate accuracy with a precision of ±0.03.		
How accurate are LLM-based synthetic inter-scale correlations (for scales consisting of at least three items)?	The synthetic estimates will exhibit an accuracy of $r = .89$ for the empirical inter-scale correlation coefficients obtained from survey data, as estimated in our Bayesian multi-membership regression model.	As above. With the resulting 1,558 scale pairs, we should be able to estimate accuracy with a precision of ±0.007.		

				and we would reinvestigate our precautions.

Note. We determined the planned precision to detect any deterioration in performance greater than .01 for item pair correlations. Because increasing the number of scales is costlier than increasing the number of items, the sensitivity for the reliability coefficients is a compromise with feasibility.

Data availability

We have shared all key materials on the Open Science Framework at https://osf.io/z47qs/. The existing data used for training and in the pilot study has been openly shared, we link to the original sources. We will also openly share our collected data.

Code availability

We have shared the training and analysis code on the Open Science Framework at https://osf.io/z47qs/.

Acknowledgements

The research is funded by the German Research Foundation grant #464488178 to Ruben C. Arslan. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank Stefan Schmukle, Anne Scheel, Julia Rohrer, Malte Elson, Taym Alsalti, Ian Hussey, Saloni Dattani, David Condon, Dirk Wulff and Jan Arnulf for helpful discussions. We also thank Jan-Paul Ries, Lorenz Oehler, and Sarah Lennartz for comments on an earlier version of this manuscript.

Author contributions

B.E.H.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Visualization, Writing - original draft, and Writing - review & editing.

R.C.A.: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Visualization, Writing - original draft, and Writing - review & editing.

Competing interests

The authors declare no competing interests.

Results & Discussion

To follow – this is a draft of a Stage 1 Registered Report before peer reviews.

References

Abdurahman, S., Vu, H., Zou, W., Ungar, L., & Bhatia, S. (2024). A deep learning approach to personality assessment: Generalizing across items and expanding the reach of survey-based research. *Journal of Personality and Social Psychology*, *126*(2), 312–331. https://doi.org/10.1037/pspp0000480

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework* (No. arXiv:1907.10902). arXiv. https://doi.org/10.48550/arXiv.1907.10902

Anvari, F., Alsalti, T., Oehler, L., Hussey, I., Elson, M., & Arslan, R. C. (2024). *A fragmented field: Construct and measure proliferation in psychology*. https://doi.org/10.31234/osf.io/b4muj

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Bong, C. H. (2014). Predicting Survey Responses: How and Why Semantics Shape Survey Statistics on Organizational Behaviour. *PLoS ONE*, *9*(9), e106361. https://doi.org/10.1371/journal.pone.0106361

Bainbridge, T. F., Ludeke, S. G., & Smillie, L. D. (2022). Evaluating the Big Five as an organizing framework for commonly used psychological trait scales. *Journal of Personality and Social Psychology*, *122*(4), 749–777. https://doi.org/10.1037/pspp0000395

Berufsverband Deutscher Psychologinnen und Psychologen. (2022). *Ethische Richtlinien der Deutschen Gesellschaft für Psychologie e. V. und des Berufsverbandes Deutscher Psychologinnen und Psychologen e. V.[Ethical guidelines of the German Society for Psychology and the Professional Association of German Psychologists]*. DGPs. https://www.dgps.de/die-dgps/aufgaben-und-ziele/berufsethische-richtlinien/

Condon, D. M. (2017). *The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model*. PsyArXiv. https://doi.org/10.31234/osf.io/sc4p9

Condon, D. M., & Revelle, W. (2015). Selected Personality Data from the SAPA-Project: On the Structure of Phrased Self-Report Items. *Journal of Open Psychology Data*, *3*. https://doi.org/10.5334/jopd.al

Condon, D. M., Roney, E., & Revelle, W. (2017). A SAPA Project Update: On the Structure of phrased Self-Report Personality Items. *Journal of Open Psychology Data*, *5*(1), Article 1. https://doi.org/10.5334/jopd.32

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113*(3), 492–511. https://doi.org/10.1037/pspp0000102

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Cutler, A., & Condon, D. M. (2022). Deep lexical hypothesis: Identifying personality structure in natural language. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000443

Digman, J. M. (1990). Personality Structure: Emergence of the Five-Factor Model. *Annual Review of Psychology*, *41*(1), 417–440. https://doi.org/10.1146/annurev.ps.41.020190.002221

Elson, M., Hussey, I., Alsalti, T., & Arslan, R. C. (2023). Psychological measures aren't toothbrushes. *Communications Psychology*, *1*(1), Article 1. https://doi.org/10.1038/s44271-023-00026-9

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Fyffe, S., Lee, P., & Kaplan, S. (2024). "Transforming" Personality Scale Development: Illustrating the Potential of State-of-the-Art Natural Language Processing. *Organizational Research Methods*, *27*(2), 265–300. https://doi.org/10.1177/10944281231155771

Greene, R., Sanders, T., Weng, L., & Neelakantan, A. (2022). New and improved embedding model. *Open AI Blog*. https://openai.com/blog/new-and-improved-embedding-model

Guenole, N., D'Urso, E. D., Samo, A., & Sun, T. (2024). *Pseudo Factor Analysis of Language Embedding Similarity Matrices: New Ways to Model Latent Constructs*. OSF. https://doi.org/10.31234/osf.io/vf3se

Hernandez, I., & Nie, W. (2023). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*, *76*(4), 1011–1035. https://doi.org/10.1111/peps.12543

Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-Based Deep Neural Language Modeling for Construct-Specific Automatic Item Generation. *Psychometrika*, *87*(2), 749–772. https://doi.org/10.1007/s11336-021-09823-9

Hugging Face model hub. (n.d.). Retrieved June 2, 2023, from https://huggingface.co/sentence-transformers/all-mpnet-base-v2

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

Hussey, I., Alsalti, T., Bosco, F., Elson, M., & Arslan, R. C. (2024). *An aberrant abundance of Cronbach's alpha values at .70*. https://doi.org/10.31234/osf.io/dm8xn

Kjell, O. N. E., Kjell, K., Garcia, D., & Sikström, S. (2019). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods*, *24*(1), 92.

Kjell, O. N. E., Kjell, K., & Schwartz, H. A. (2024). Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Research*, *333*, 115667. https://doi.org/10.1016/j.psychres.2023.115667

Kopalle, P. K., & Lehmann, D. R. (1997). Alpha Inflation? The Impact of Eliminating Scale Items on Cronbach's Alpha. *Organizational Behavior and Human Decision Processes*, *70*(3), 189–197. https://doi.org/10.1006/obhd.1997.2702

Larsen, K. R., & Bong, C. H. (2016). A Tool for Addressing Construct Identity in Literature Reviews and Meta-Analyses. *MIS Quarterly*, *40*(3), 529–551. https://doi.org/10.25300/MISQ/2016/40.3.01

NVIDIA, Vingelmann, P., & Fitzek, F. H. P. (2022). *CUDA, release: 11.7.1*. https://developer.nvidia.com/cuda-toolkit

Pretrained Models—Sentence-Transformers documentation. (n.d.). Retrieved March 5, 2024, from https://www.sbert.net/docs/pretrained_models.html

R Core Team. (2023). *R: A Language and Environment for Statistical Computing* (Version 4.3.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (No. arXiv:1908.10084). arXiv. https://doi.org/10.48550/arXiv.1908.10084

Rosenbusch, H., Wanders, F., & Pit, I. L. (2020). The Semantic Scale Network: An online tool to detect semantic overlap of psychological scales and prevent scale redundancies. *Psychological Methods*, *25*(3), 380–392. https://doi.org/10/gg5rn7

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 815–823. https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Schroff_FaceNet_A_Unified_2015_CVPR_paper.html

Sharp, C., Kaplan, R. M., & Strauman, T. J. (2023). The Use of Ontologies to Accelerate the Behavioral Sciences: Promises and Challenges. *Current Directions in Psychological Science*, *32*(5), 418–426. https://doi.org/10.1177/09637214231183917

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). *MPNet: Masked and Permuted Pre-training for Language Understanding* (No. arXiv:2004.09297). arXiv. https://doi.org/10.48550/arXiv.2004.09297

Tunstall, L., Werra, L. von, Wolf, T., & Géron, A. (2022). *Natural language processing with Transformers: Building language applications with Hugging Face* (First edition). O'Reilly.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

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

Williams, A., Nangia, N., & Bowman, S. R. (2018). *A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference* (No. arXiv:1704.05426). arXiv. https://doi.org/10.48550/arXiv.1704.05426

Wulff, D. U., & Mata, R. (2023). *Automated jingle–jangle detection: Using embeddings to tackle taxonomic incommensurability*. PsyArXiv. https://doi.org/10.31234/osf.io/9h7aw