Editorial decision

Dear Authors,

Thank you for your Stage 2 submission. Two reviewers from Stage 1 returned, and I invited an additional reviewer to replace a third reviewer from Stage 1 who wasn't able to evaluate the Stage 2 report. Overall the three reviewers commented positively on the submission, but pointed out somewhat minor but still important areas of improvement in your write-up.

I am happy to invite you to submit a revision & response letter, and anticipate a quick turnaround afterwards. In your revision/response, please address the reviewers' comments while paying attention especially to the following points:

Ensure the tense is consistent & appropriate throughout (e.g. in Intro, the validation study is now done and so you no longer "plan to validate"; "All necessary support is in place for the proposed research." is now redundant, etc.)
Include an appropriate related interests statement or similar
Overall the transparency of the document and associated code/programs deserves praise, but as noted there are some areas where you should update your code/documentation as noted by HC-P
Update discussions of Wulff & Mata and pre-trained model concerns as noted by ZH
Are you able to better link to the supplementary online materials? e.g. there are references to "Supplementary Notes" but I had to dig around in the OSF repo to find the correct file(s). I understand this may be a limitation of the OSF but perhaps you can see if this could be improved (with e.g. directly linking to files' persistent DOIs.)

Replies to Johannes Breuer, Reviewer #1

As for the stage 1 version of this paper, I have enjoyed reading the stage 2 report and believe that this can make a valuable and impacful contribution to different fields (esp. psychometrics/scale development and research on the use of AI/LLMs for science).
The authors have followed the preregistered procedures laid out in the stage 1 report. The few minor deviations (e.g., in the sampling/inclusion criteria for respondents or the sample size) are transparently reported.
I only have a couple of minor remarks that should be relatively easy to address for the authors. I will list those in chronological order in the following.
I would suggest changing the language in the Introduction (and elsewhere) describing the validation study from future tense to past tense (or present tense) as it has now been conducted.

> ***Reply by the authors****: We thank the reviewer for their positive comments on our manuscript and for their willingness to help us improve our manuscript. We have made edits to change the tense from future to past tense across the manuscript.*

On p. 18, 2nd para, the upper bound of the CI for the manifest correlation for the accuracy of synthetic scale correlations is missing. It appears that the formatting of the whole parentheses in which this is contained is not correct/broken.

> ***Reply by the authors****: This was indeed a formatting error, which we have now corrected.*

A minor note related to Figure 5: From a data viz perspective, I prefer it if the y-axis and the bars start at 0 (i.e., without any whitespace between the bars and the x-axis).

> ***Reply by the authors****: We changed Figure 5 accordingly.*

Regarding Figure 6: The Figure distinguishes between “latent outcomes (SEM)” and “manifest”. This distinction appears for the first time in this form/wording here. Hence, I think that this should be (further) explained in the paper before the figure is presented.

> ***Reply by the authors****: We agree that merely mentioning the SEM approach in the supplement while mentioning it in Figure 6 is confusing to the reader. The revised manuscript explains our use of accuracy metrics now in more detail in the analysis plan on p. 14.*
>
> *“[...] First, we used a structural equation modeling (SEM) approach where we fixed the residual variance of empirical estimates to the average sampling error variance and allowed manifest synthetic estimates to correlate with the latent variable. [...]”*

In the Discussion, maybe the parts addressing the general/broader context (and the implications of the presented findings for this) could be extended a bit to elaborate further on what the results (can) mean for research practices in psychology and other social and behavioral sciences (see, e.g., the recent paper by Binz et al., 2025).

> ***Reply by the authors****: We have widened our discussion on p. 22 a bit to integrate our perspective into a broader stance on human-AI collaboration. Both of us, the authors, remain agnostic to some of the epistemological questions raised in Binz et al. (2025) and by others in recent debates, and have a more pragmatic position on the role of AI in psychology. We have therefore focused on a concise discussion regarding the potential impact of this line of research on the scientific ecosystem.*
>
> *"We believe that this line of work exemplifies a responsible integration of LLMs into research, which is a topic of current debate (Binz et al., 2023). Specifically, the collaborative circumstances in scale development carry minimal risk for harmful effects on the scientific ecosystem. False negatives (i.e., the model fails to detect redundant scales) would merely maintain the status quo, which has led to construct proliferation in the first place. False positives (e.g., the model incorrectly flags two measures as redundant) would require researchers to verify this empirically before drawing conclusions. This balanced approach, where LLMs accelerates discovery while human researchers retain interpretive authority, should characterize a productive human-AI collaboration across the social and behavioural sciences."*

Literature cited in this review

Binz, M., Alaniz, S., Roskies, A., Aczel, B., Bergstrom, C. T., Allen, C., Schad, D., Wulff, D., West, J. D., Zhang, Q., Shiffrin, R. M., Gershman, S. J., Popov, V., Bender, E. M., Marelli, M., Botvinick, M. M., Akata, Z., & Schulz, E. (2025). How should the advancement of large language models affect the practice of science? Proceedings of the National Academy of Sciences, 122(5), e2401227121. https://doi.org/10.1073/pnas.2401227121

Replies to Hu Chuan-Peng, Reviewer #2

I am pleased to review this Stage 2 Registered Report. Overall, the work is rigorous, transparently documented, and addresses its objectives effectively. The results demonstrate the model's ability to mirror real-world reliability coefficients, scale correlations, and covariance patterns. Please see below for my revision suggestions to strengthen the manuscript further:

(1) Analytical reproducibility concerns
The GitHub repository (https://github.com/synth-science/surveybot3000), which was linked to OSF, requires updates to align with the Stage 2 submission. For example, the repository's README files appear outdated and do not reflect the current workflow or documentation; Critical files/folders necessary for full reproducibility are missing (e.g., ../synth-rep-dataset/, ignore/, docs/, ignore.random_scales_rr.rds). If these omissions are intentional (e.g., due to privacy constraints), this should be explicitly justified in the repository and manuscript.

> ***Reply by the authors****: We thank the reviewer for alerting us that work needs to be done for full reproducibility. We initially gitignored several files so as not to clutter the repository while it was work in progress. However, all data necessary for full reproducibility can be shared and now are. We have updated the READMEs and shared anonymised data from which the analyses can be produced. We have also shared intermediate files (such as random_scales_rr.rds) which should be in principle reproducible given that we set a random seed, but better safe than sorry.*

(2) Clarify how accuracy was estimated in methods
The manuscript should briefly summarize how prediction accuracy was quantified (e.g., error metrics, validation procedures). While Supplementary Note 1 mentions two approaches, the main text lacks sufficient detail for readers to evaluate methodological rigor. Please clarify this in the Methods section.

> ***Reply by the authors****: We agree that this wasn't sufficiently addressed in the main manuscript, and other reviewers have raised similar points. In the revised version of the manuscript, we have elaborated on this and now included a more thorough description of the accuracy metrics employed in our research on p. 14.*
>
> *"The central performance metric in this study is accuracy, defined as the convergence between synthetic and empirical estimates (not to be conflated with evaluation metrics of binary classifiers). We thus refer to manifest accuracy as the Pearson correlation between synthetic and empirical coefficients. We quantified latent accuracy using two complementary approaches that account for sampling error in empirical estimates. [...]"*

(3) Discussion of figure 4
The variation in prediction error across prediction types (Figure 4) is intriguing but not mentioned in the discussion. Could these patterns reflect meaningful differences in model performance? Consider addressing this briefly in Discussion.

> ***Reply by the authors**: Figure 4 holds important information that is central to understanding the limitations of our model. Although we have discussed the implications of the relationship between prediction error and synthetic estimates on p. 21, we did not explicitly reference the figure in that section, which may make it difficult for readers to link the passage to the figure. We have now corrected this.*

(4) Conflict of interest statement
A conflict of interest statement is absent. Given the first author's affiliation with magnolia psychometrics GmbH (a commercial entity), please confirm whether this paper's recommendations or outcomes could be perceived as benefiting their commercial interests.

> ***Reply by the authors**: Although we as authors see no direct conflict of interest with this research, we agree that BEH's affiliation to a private company merits a more explicit disclosure. We have updated the "Competing interests" section of our manuscript as follows:*
>
> *"Björn E. Hommel is affiliated with magnolia psychometrics GmbH, a private consulting agency which has agreed to maintain the app accompanying this paper, currently hosted on Hugging Face. There are no competing interests, financial or otherwise, related to this research. The SurveyBot3000 model is freely licensed under Apache 2.0, so that both non-commercial and commercial future applications may be developed."*

Replies to Zak Hussain, Reviewer 3

An important and timely contribution - I look forward to seeing this work published! I would, however, suggest the following (minor) revisions:

I would clarify in the abstract which 'out-of-sample accuracy' metric is being used (Pearson correlation, coefficient of determination etc.).

> ***Reply by the authors****: Thank you for your thoughtful comments that have helped us to refine our manuscript. We've revised the abstract and briefly clarified our general approach to measuring model performance.*

Perhaps consider re-wording 'out-of-sample accuracy' to 'out-of-sample performance'. The term 'accuracy' often suggests that the task being performed is a classification task, which is not the case here.

> ***Reply by the authors****: We agree and understand that the term "accuracy" is much more narrowly defined in the ML literature and exclusively used for performance metrics in classification tasks. The term is more loosely employed throughout behavioural research (we consider behavioural researchers our main audience). The added clarification in our abstract and an additional passage on p. 14 on how we operationalize accuracy should hopefully dispel any ambiguities.*
>
> *"The central performance metric in this study is accuracy, defined as the convergence between synthetic and empirical estimates (not to be conflated with evaluation metrics of binary classifiers). We thus refer to manifest accuracy as the Pearson correlation between synthetic and empirical coefficients. We quantified latent accuracy using two complementary approaches that account for sampling error in empirical estimates. [...]"*

"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."

This paragraph appears to blur the distinction between related but distinct concepts. In the context of language modeling, fine-tuning is typically understood as updating at least some model weights for a specific task. The phrase "the model then retains its originally learned weights" could therefore be misleading unless referring to methods where weights remain frozen (e.g., in-context learning, of which few-shot learning is a special case). If the authors intended to describe an approach where the weights remain frozen, I would suggest using "in-context learning" rather than "fine-tuning". However, given that the research later performs full-model fine-tuning (i.e., with weight updates), I assume this is what the authors mean. In that case, it would be helpful to specify the distinction more explicitly to avoid conflating fine-tuning

with few-shot learning, which often relies on frozen weights and contextual adaptation rather than explicit weight modification (see, e.g., Brown et al., 2020). In this case, I would recommend sticking to the term "transfer learning" instead of "few-shot learning".

> *Reply by the authors: We absolutely agree that the phrasing in this paragraph (p. 4) was off and misleading. We rephrased the passage about fine-tuning, which indeed could be interpreted to suggest as if all model weights are frozen during fine-tuning. We also agree that the term "few-shot learning" nowadays (since the GPT3 paper) is more frequently employed when describing in-context learning and therefore inadequate.*
>
> *"Fine-tuning often involves slight architectural adjustments to the model's output layer, although the term is used somewhat inconsistently in the literature to describe various adaptation approaches."*

"However, their [Wulff & Mata, 2023] 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."

Although earlier version of Wulff & Mata's work do not adapt their model to the domain of the survey items, the most recent versions does (Wulff & Mata, 2025).

> ***Reply by the authors****: As our manuscript was submitted as a two-staged registered report, we adhere to PCI-RR's submission requirements which state that "[...] introduction, rationale and stated hypotheses [should remain] the same as the approved Stage 1 submission". We could of course update the reference, but this would necessitate further edits because it is no longer in the timeline in which the project developed. Nevertheless, we of course agree that the final update Wulff & Mata needs to be referenced. The revised manuscript therefore now properly cites and discusses Wulff & Mata (2025) in the discussion (p. 21). Similarly, we also cite the published version of Schoenegger et al. (2025) which includes a (difficult to interpret) comparison between PersonalityMap and SurveyBot3000.*
>
> *"In a recent update to their work, Wulff & Mata (2025) have adopted fine-tuning techniques that improve upon their earlier results, yielding accuracies that approach the performance we report here, but limited to absolute correlations."*

As a non-expert in psychometrics, it is not immediately clear to me what the value is of being able to predict polarity/negative correlations. Perhaps the value of doing this could be made clearer somewhere early in the paper, especially since it appears to be a key contribution of the work. In particular, it would be useful to have an explanation of how predicting polarity might help researchers 'evaluate new measures against existing scales, reduce redundancy in measurement, and work towards a more unified behavioural science taxonomy', as expressed in the abstract as a main contribution of the work.

***Reply by the authors**: A substantial portion (41%) of pairwise item-correlations in our validation study are negative. These negative correlations represent meaningful psychological relationships. Merely predicting the magnitude of correlations while ignoring their sign drastically reduces model performance. This limitation is visible in the croissant-shaped plots in Figures 2 and 3 for the SBERT model, which struggles with negative correlations. Our calibration step addresses this limitation. When assessing whether two constructs overlap, determining if they correlate positively or negatively is just as important as the correlation's strength. Similarly, reliability could not be approximated well using synthetic estimates if unintended negative loadings cannot be detected.*

As noted in our discussion (p. 23), even modest improvements in sign prediction yield substantial gains in accuracy: "If we imagine that a human user of our app can correct the coefficient sign in these small-scale applications, this would improve manifest accuracy by .11, yielding an overall convergence of .68 between synthetic estimates and empirical correlations."

We've now clarified this contribution on p. 21: "A novel calibration step further enables the model to predict negative correlations (e.g., opposing items), more accurately reflecting the empirical distribution of coefficients."

In general, I would find it helpful if the authors expanded on the main differences/contributions of their work relative to Wulff & Mata (2025). I appreciate that these two pieces of work were carried out in parallel, and thus some overlap is inevitable. However, I get the impression that differences/contributions of the present work could be emphasized more clearly.

***Reply by the authors**: We have not explicitly contrasted our work against Wulff & Mata (2025), but against several similar publications referenced in the paragraph on p. 21. (i.e., Hernandez & Nie, 2023; Schoenegger et al., 2024; Wulff & Mata, 2023; Wulff & Mata, 2025). The four points mentioned in the subsequent passages are what we consider the distinguishing contributions of our work.*

"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."

This is an important consideration. However, it is not obvious to me that the pre-trained model used (all-mpnet-base-v2) is free from such concerns. Perhaps the authors could expand on the extent to which they believe information on their survey items might have leaked into the datasets (Wikipedia, BooksCorpus, OpenWebText, CC-News, and Stories) used to pre-train all-mpnet-base-v2.

> ***Reply by the authors****: We agree that this is not self-evident. We mostly sought to address this concern by focusing on recent scales for the validation study. However, because we also tried to optimize for content breadth, we ended up including some older scales, which could have been part of pre-training data.*
>
> *We therefore conducted an extensive search in the corpora which were used to train all-mpnet-base-v2. Specifically, we searched for substrings in ~ 90 GB of text data which matched any of the item text used in our pilot or validation study. Roughly one third of items had matching substrings in the training data. In many cases, these were short substrings like "I was happy" which occurred in natural language outside of a survey context. For the validation study, we investigated matches in greater depth and found that items for two instruments were part of the training corpus (simply as part of a survey description, not along with quantitative data on correlations). However, excluding them had little impact on accuracy. In the revised manuscript, we describe this in more detail in Supplemental Note 11, as now referenced on p. 5.*

"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."

It may not be entirely accurate to say that the model's predictions are based **only** on the **semantic information** contained in the items. The authors already noted that pre-trained models could be leveraging other sorts of information to make their predictions, such as co-occurence patterns of the items found "in online studies and public item repositories". Even smaller language models like all-mpnet-base-v2 likely encode more than just word meanings. For instance, such models have been shown to predict response times and reading difficulty, which can be impacted by other factors such as item complexity, social desirability, or phonetics.

> ***Reply by the authors****: There is an interesting debate to be had here. In linguistics, the distributional hypothesis states that meaning is distilled into tokens by regarding their context. In that sense, it is our view that there is nothing but semantic information that can be regarded by an agent. It seems arbitrary to distinguish between semantic meaning derived from contextual statistics and meaning derived from item co-occurrence patterns. Then again, following this logic would also challenge other traditionally useful distinctions, such as the boundary between semantics and pragmatics. Hoping that a minor edit to a passage from Stage 1 of our RR is permissible at this Stage, we've clarified this sentence to read:*
>
> *"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* ***(i.e., adopting the notion of distributional semantics which considers all contextual patterns as inherently semantic)****."*

In general, more thorough discussion of what the model could be leveraging to predict item similarity would benefit this work, especially since it has implications for a possible upper bound on model performance when predicting item correlations. This could also be informed by speculation/research on the extent to which observed item correlations reflect semantic similarity versus other factors.

> ***Reply by the authors****: This is important to future work, but we would like to avoid speculating on it in this particular work of ours, because the few predictions we felt equipped to test yielded equivocal results. However, we have hinted at the direction where future research should be headed to answer these questions on p. 23. To clarify our point, we have extended the paragraph as follows:*
>
> *"Instead of comparing vectors monolithically, future approaches could isolate psychometrically relevant information by separating residual features in vector space.* ***This decomposition approach may help establish theoretical upper bounds on prediction accuracy by distinguishing between different types of semantic content captured in vector space, including conceptual meaning, but also peripheral semantic information such as survey response tendencies.****"*