Editorial decision

Dear Dr Hommel & Dr Arslan,

Thank you for submitting your Stage 1 Registered Report "Language models accurately infer correlations between psychological items and scales from text alone" for evaluation at PCI: RR. Thank you also to the three reviewers who submitted their evaluations. I concur with the reviewers that the manuscript is interesting and you have already done exciting work with the pilot study; yet there are aspects of the proposed study that need additional work. I am therefore inviting you to submit a revised manuscript with responses to the reviewer's comments.

I would suggest you pay particular attention to ensuring data quality, and further clarify and justify the US-based representative sampling plan (as suggested by JB). An anonymous reviewer also suggested sticking closer to a (psych) traditional intro, study 1-methods-findings,... structure to guide readers. Finally if hypotheses must be made, HC-P suggests that they should be further clarified.

Thank you, and looking forward to the revision
Matti Vuorre

Reply to Dr. Matti Vuorre

Dear Dr. Vuorre,

Thank you for your letter and the invitation to submit a revised version of our Stage 1 Registered Report, entitled "Language models accurately infer correlations between psychological items and scales from text alone." We are grateful for the insightful feedback from you and the three reviewers, which has significantly contributed to improving our manuscript.

In this revised version, we have carefully addressed the reviewers' specific comments, as well as your own suggestions. Notably, we have: a) Strengthened our data quality measures, incorporating recommendations from Goldammer et al. (2020) and Yentes (2020) to identify and exclude participants with problematic response patterns. b) Clarified our sampling plan, specifying the use of Prolific's quota sampling to balance age, sex, and ethnicity based on the 2021 US Census data. c) Further clarified our hypotheses and the quantitative criteria for evaluating them, as suggested by Reviewer 3.

We believe we have successfully addressed all comments and concerns raised by the reviewers, as outlined in our detailed responses below. Additionally, we have expanded the

online supplemental materials, which now provide further details on the training procedure of the SurveyBot300.

We confirm that no data generated as part of this manuscript are currently under review or published elsewhere. Furthermore, this manuscript is original, has not been previously published, and is not under concurrent consideration elsewhere. All research conducted was in strict adherence to appropriate ethical guidelines.

We hope that the revised manuscript meets your and the reviewers' expectations, and we look forward to your feedback.

Sincerely,

Dr. Björn E. Hommel & Dr. Ruben C. Arslan

Replies to Reviewer 1

Overall, I find this paper to be an important and useful step forward in NLP modeling of psychological items and scales. As always, there is some room for improvement on the margins, but I think it's a great sign that I am eagerly awaiting the results of this Registered Report.

> *We thank the reviewer for their positive assessment.*

In terms of feedback, it seems important to mention that my comments below are not ordered in terms of priority. They match the ordering of the Registered Report instead.

Introduction:

I think the first 2-3 pages try to cover too much. Especially the 2nd page (roughly paragraphs 5-8). As a reader, I felt that it took a lot of effort to get through, though I was also aware that the

authors were trying to contextually frame NLP/AI modeling for psychologists who may be encountering it for the first time (not easy to do!). To be clear, the authors did a nice job with the text that's given -- I view it as accurate and reasonably concise (though it is unavoidably jargony). But, I think much/all of this could be "off-loaded" by directing readers to prior work using citations that the the authors already have in place. This technique is done well later when explaining bi-directional encoding. In fact, I think it would (almost) be okay to simply cut the first ~1000 words, beginning the paper with the last paragraph of the 2nd page ("Wulff & Mata (2023) used..."), then expanding that paragraph extensively. In other words, I'd like to hear more about negation (a really big issue!), why scale aggregation is better, why ada-002 should not be seen as the gold standard, why training a model for the task at hand is a big deal (it is!), etc. As for the consequences of cutting all the background info, I think this is probably okay. Given the typical timing of registered reports, many readers will be caught up on these techniques (and their relevance to questions of structure) by the time the paper comes out. For those that are not yet familiar, something other than a technical empirical paper like this would probably be needed (e.g., a "Current Directions" article).

> *We thank the reviewer for their advice on structuring the manuscript. We are happy to expand as requested. But we decided not to take the recommendation to just send readers elsewhere for an introduction for the following reasons. First, the manuscript is already out as a preprint, so to us, it does not just matter what literature will be available when the final manuscript is published. Second, we read some of the existing introductions to the topic in our internal journal club. The consensus was that these were very jargony and some of our club members who do not work with LLMs quickly felt lost. Third, all introductory papers that we would consider citing are still preprints and might change considerably. Our didactic approach here was to use the jargon many psychologists already know from their basic training about measurement, learning, and factor analysis and bridge it with the LLM jargon. We have attempted to reduce jargon or to explain it where necessary in this revision. Fourth, in this quickly developing field, we think it is important to cite important prior work, but also briefly outline how it differs from our approach in focus and approach, because the differences between papers working with pre-trained models and fine-tuned models etc. are not apparent at a glance to lay readers of the literature.*

Personally, I would not call these synthetic correlations. For me, "synthetic correlation" already has already has a meaning (though I acknowledge that this term is not widely used, especially

in social science). ... With sufficient cooccurrences among two variables, it's possible to derive plain-old, empirically-observed correlations; with zero-or-too-few cooccurrences, it is sometimes possible to derive *synthetic* correlations based on other known parameters (like, by estimating distributions based on joint cooccurrences with other variables). I don't think this meaning matches the coefficients called 'synthetic' in the paper, but I may be mistaken. Either way, I think it's more important and informative to reference these correlations as being *semantic*, as this would help to emphasize the difference from typical (ratings-based) corrs used in psych research. I realize this is a major change of a small thing bc the term 'synthetic' is in many locations. But, I really think it's worth considering. They are semantic (which is important to highlight), and I don't think they're synthetic (but maybe... if so, explain how?).

> *As the reviewer has noted, this may be a question of preference and background. Of course, we deliberated a while on what to call these quantities. Our reason not to use "semantic" was that the word is overloaded as well and produced unfortunate associations (in our early readers) that the model only understands the meaning of words as defined in a dictionary. But our model has a world model because it has learned that items "I love to chat" and "I have high stamina" are correlated even though the dictionary meaning is not closely related.*
>
> *Synthetic **data** is the only widely used term including the word synthetic in psychology that we are aware of and denotes generating data to resemble empirical data. We think this fits. We favour synthetic because it has an implied hint of caution that it is not "the real thing" and contrasts well with empirical. It also makes it clear that the data were generated by an artificial process, which semantic does not achieve (it could be understood as similarity rated by humans too). We think as long as we clearly define how we use the term the potential for confusion with concepts that will be unfamiliar to most readers anyway is low. To this end, we have rephrased the text where we introduce the term as follows:*
>
> *The distances between item pairs in vector space produce <u>what we will call</u> synthetic item correlations*

Stage 1: Polarity Calibration. The description is wonderful! But, I was left with many questions and wanting more. "The fine-tuned model was then trained to predict this new criterion, which

combined the magnitude and direction of the similarity." Can you say more about how this was done, and (more importantly) how it went? I thought Figure 2 might address the latter question to some extent, but the caption was largely inadequate for making sense of the images, in my opinion. Of course, this was particularly true for my first pass through the manuscript as I did not yet know where the empirical data came from (it's introduced in the next section). I don't think a lot of detail is needed about fine-tuning, but a few more sentences would help. This will be completely new for most readers (even many who know a bit about AI).

> *We agree that our description of this particular fine-tuning stage should cover additional details, as polarity calibration was a crucial methodological decision. In the manuscript, we now elaborate as follows:*
>
> *"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)."*
>
> *Details on the effectiveness of this training step can now be found in Supplementary Note 6. Specifically, we compare two versions of the SurveyBot3000 and subsequently demonstrate, that omitting the polarity calibration stage significantly impacts the model's ability to correctly identify opposing sentences.*
>
> *We believe these additions address the reviewer's concerns and provide readers with a clearer understanding of why this training approach is essential in teaching the model to correctly predict the relationship in item pairs containing reversed items.*

Stage 2... Again, the text provided is very good (i.e., the description of data sources and partitioning), but some further description of how the authors "fine-tuned the model to focus on text segments that convey psychologically relevant information" is missing. I also think some readers may think this interpretation is sort of a stretch... that it sort of assumes the semantic correlations should be roughly equivalent to the rating data. (I actually suspect this may prove to be the case, but I also think it should remain an open question for a while -- certainly, I can

argue that they aren't the same, that psychology is not just language...). At the end of this section, the comment that the focus was "on enhancing the model's ability in predicting item correlations within the test partition" seems more straight-forward and accurate.

> *We rewrote this section to more clearly say that "focusing on text segments that…" was a result of the goal function we implemented (predict item correlations), not the mechanism by which we implemented it. Generally speaking, we would say our second training step was meant to teach the model about the structure of human personality, as captured in self-report items. So, we do not think our model only learned about language in the dictionary sense. It had access to empirical data about correlations between items. That it performs well in holdout data will be partially because some items are closely semantically related to others and partially because personality structure is somewhat sparse, so new item relationships can be predicted if you know about old item relationships.*

Why "mnemosyne"? It's cool but this may warrant some description if the hope is that the name will catch on or be used beyond the current project. If that's not the plan, it may be easier to give it a more informative name (boring is sometimes functional).

> *We changed mnemosyne to "a curated corpus of item texts and correlations between items".*

In the pilot study, using random subsets of items to evaluate reliability estimates that are not biased seemed strange to me. I think this should probably be dropped. The analysis of reliability estimates for published scales is useful and interesting, but randomly subsetting items and then using some exclusion criteria to estimate de novo reliabilities is not providing additional information beyond what is shown with the full item-level correlation results in my opinion. If anything, I find it a bit misleading.

> *We do not intend to mislead readers, hence the transparent reporting. Possibly, the reviewer thought we were "hiding" the results without the newly formed scales. They were reported in the supplement. We now report them in the manuscript. We describe the process thusly:*
> *"Empirical Cronbach's alpha values had a mean of .76 (SD = .11) 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 think the first two sentences are trivially true. The published literature (and certainly Bainbridge et al.) mainly contains scales that reached the cutoff of .70 (often attributed to Nunnally), but in scale development scales often fall short of this criterion, which is of course one reason why we do scale development. Of course, it would be possible to simulate the development of scales more realistically than what we did. We considered investing more effort here somewhat peripheral because the main quantity of interest are the item-level results, but we disagree that there is no additional information beyond the full item-level correlations. Perhaps someone who is steeped in psychometrics might be able to infer from the item-level results the intuition that reliability predictions get less accurate at lower levels (because when items are barely correlated, errors about direction, not just magnitude are more likely). But those who cannot do so surely benefit from seeing a scatter plot with a prediction interval, which are very easy to understand. We now report all of these results in the main text. We had previously reported the performance for real scales only in the supplement (r = .63). For what it's worth, the value of .86 for the set including the random scales is close to what we get when we disattenuate .63 for a restriction from SD=.235 to SD=.099 ($r_{unrestricted}$ = .89) using psychometric:cRRr.

Revised text:

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 variance in the real scales (SD = .10 compared to SD = .23 in the combined set).

For the scale analyses, I think it would be useful to include a detailed analysis of cases where there were very high and very low r's (especially low). This might provide some ideas about what topics aren't explained by semantics alone. These results seem especially important for the pre-registered work to be done next.

> *We are not sure we correctly understood this point. The interactive plot in our supplement allows the reader to access all pairs of empirical/synthetic correlations. On average, estimated synthetic rs are low when empirical rs are low and vice versa. Deviations from the regression line occur, but the scatter around the regression line mixes some aspects (smaller absolute correlations have larger standard errors, and scales with many items are more likely to have high manifest correlations with related scales by virtue of their high reliability). We think these factors explain why some dots fall outside the yellow prediction interval in the plot. We did not see any further signal to extract from these patterns, but we think this exercise might be worthwhile to report as an exploratory result about generalizability after we collect new data.*
> https://rubenarslan.github.io/surveybot3000/#Interactive_plot14

In the Measures section, I think it's imperative to mention an inclusion criteria relating to IP. I would recommend making this explicit and omitting any content that has not been clearly licensed/put in the public domain. This is actually a major issue, in my opinion, and it's much better to leave out any measures with unclear IP status. If APA PsycTests has done this for you (I think they have, right?) -- that's great, it will make it super easy to add a statement in the inclusion criteria to put reader's concerns to rest. Of course, not everything that's publicly available for various uses is public domain. In my experience, this can get quite complicated!

> *Reviewer 2 also brought up the point of proprietary measures. We verified all planned measures permit sharing and using the items openly and make this explicit in the Measures section (see p. 13).*

Also in Measures: I agree that a uniform response pattern is best, but I would give slight preference to a 6 point response format. I am aware that the prior work here suggests that 5 vs 6 is not a critical point, but I have some good reasons. For one, psychometrically, there is

evidence than >5 is better than 5 or less (6 seems to be the threshold where most things do fine when treated as continuous), and that having an even number of options is slightly better or equivalent to odd (probably better bc no neutral option). Second, and more important, I am also familiar with a very large dataset including many of these measures that uses a 6 point response scale, and it would be nice if the data sets could someday be directly compared (even descriptively).

> *We changed the Likert scale from 7 to 6 point based on this advice.*

The note for Table 1 has an error: "Because increasing the number of scales is costlier than number of scales..."

> Corrected: 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.

Finally, an overall comment: I have some confusion about the structure of this Registered Report. Every outlet is different, but my own experiences with this submission format have still made use of a familiar Intro/Method/Result/Discussion structure. This manuscript departs from that structure quite a lot, which is manageable but a bit confusing. I guess my point here is just to suggest that the authors take care to order the full/complete manuscript in a way that's easy for readers to digest.

> *Our stage 2 Registered Report will make use of this structure, so the Results and Discussion for the main study are currently missing. The final article will have the familiar structure.*

Replies to Johannes Breuer, Reviewer 2

First of all, I want to say that I sincerely enjoyed reading the manuscript. It covers a highly relevant and timely topic and I am quite certain that the study can make a very valuable contribution to the field. The manuscript is also written in an engaging and accessible manner, which is especially relevant for a fairly technical and sophisticated methodological study such as this one. I also very much appreciate the provision of the ample and helpful supplementary materials (code & data on OSF & GitHub; statistical reports and interactive plots via website + app on huggingface). As a sidenote on this: I tested the app on huggingface with a scale from my area of research and it worked quite well.

> *We thank you for the positive comments and are glad to hear the app worked in your area.*

That being said, I have a few (mostly minor) questions and suggestions that the authors might want to consider a) for the preregistration and the design of the study or b) for writing a full paper later on. I have divided my remarks accordingly in the following.

a) Preregistration and study design

The authors might want to consider additional data quality checks (besides attention check items and completion time) as recent research has shown that LLM use is fairly common among crowdworkers (Veselovsky et al., 2023). This could, e.g., be checks for patterns in the responses or contradictory responses to items that should be (positively) related (or, vice versa, same responses to items that should not be related).

> *Veselovsky et al. found that crowdworkers use LLMs for text production. We have not heard of crowdworkers using LLMs for answering survey questions and it would seem like breaking a fly on the wheel, i.e. the tool is overpowered for a task that a random click script will do well enough. Of course, if this truly were commonplace that would make our validation study problematically circular. However, from the experiences we have seen with assigning personality tests to LLMs, we do not think contradictory responses or response patterns would necessarily identify LLM use. In that case, going to a non-crowdworker sample would seem better. Due to our location in Germany, this would be difficult with the English-speaking SurveyBot. Because the risk seems remote at present, we will only discuss this as a potential limitation. In the unexpected case, that the SurveyBot outperforms its performance on the (pre-widely available LLM era) Bainbridge data, we would have to reassess this risk.*

The "Measures" section states that the study will also include scales “from other social sciences”. While it makes sense to assess generalizability beyond psychology, maybe it may be better to work with smaller steps and focus on psychology first, and then separately (and maybe also more systematically) assess generalizability to other disciplines/areas within the social and behavioral sciences? It may also make a difference whether a scale assesses (personality) attributes or attitudes. On a related note, the authors may want to also consider the domain of the scales in the analyses (e.g., to compare prediction accuracies for different domains).

> *We thank the reviewer for the suggestion. We have now classified the measures in our Table S5 into the five broad categories of personality, social, occupational, clinical psychology, and non-psychological attitudes and will investigate in a robustness check whether performance differs by category. We think this classification helps to show that we have mainly focused on psychology, but are testing the waters in other social sciences with our attitude scales.*

In the “Sampling Plan” section, the authors report that they plan to “collect a representative US sample of n = 450” (note: in Table 1, row 1, column 3, it says N = 400). Regardless of my personal issue with the term and concept of representativeness, it would be helpful to indicate what “representative” refers to here (e.g., distribution of age, gender, etc. reflecting the respective distributions in the 2022 US Census data. Also, I assume that this is a non-probability sample, which is a detail that should be reported explicitly, I would say.

> *The 450/400 discrepancy is explained in the second paragraph of the sampling plan, we estimate that we will have to exclude circa 50 participants. You are correct to note your reservations about the somewhat empty term "representative" here. We have rephrased this to say "a quota sample balanced for age, sex and ethnicity according to the 2021 US Census using Prolific's inbuilt quota sampling". “Representative” is the word Prolific uses, but it is quota sampling from their pool of crowdworkers, not probability sampling and we had no intention to mislead readers about this. We cannot afford a probability sample for the number of items we plan to run and most psychological studies, including those in our training data, are run on convenience and/or crowdworker samples. We think this is suboptimal for the field, but for testing the validity of the SurveyBot, we would rather not deviate from current practice too much.*

Is the exclusion of participants “who fail at least three out of five attention checks” based on recommendations from the literature or previous research by the authors? Relatedly: What is the threshold of a minimum survey completion time of 11 minutes based on?

> *We have given more thought to our attention checks and quality control measures. The revised section reads:*
>
> *“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 ≥ -.40), e) even-odd-index (≥ .20 SD above mean).”*

Regarding the hypotheses as presented in Table 1: Are the hypotheses really meant to test exact point estimates (e.g., an accuracy of exactly r = .71 for inter-item correlations as in the 1st hypothesis)? This would then mean that finding an accuracy of .7 would not confirm this hypothesis. Hence, I was wondering whether it may be more feasible to test for a certain range, such as the 95%-CIs reported in the "Introduction" section.

> *Because the 95% CIs in your pilot study, were so narrow and we plan for them to also be narrow in our validation study, we should be able to detect performance decreasing even by 0.01 (for item pairs and scale pairs). We would discuss performance that is only slightly attenuated differently than performance which is greatly attenuated. We had roughly outlined this in our Design Table but have expanded on this in response to R3.2.*

b) Full paper

The authors discuss several application possibilities for their work. Another application area where this work can be useful are simulation studies (e.g., also for a-priori power analyses) for which the correlation and reliability estimates produced by the model could be used.

> *Thank you for this great suggestion, which we incorporated as follows:*

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.

Could expert (or simply respondent) ratings or prediction markets servr as a further benchmark for assessing the model estimates?

Yes, they could. There is older work on this, which we plan to cite in the discussion (e.g., Epstein & Teraspulsky, 1986). Schoenegger et al. (2024) recently published data on the performance of lay and expert human judges as well as aggregated human judges as benchmarks. We will cite their benchmarks in the discussion, but we will not repeat the exercise here. We eschew the expense of having thousands of item pairs rated for their similarity by human judges, because even if human judges outperformed the SurveyBot, the cost human judges command would still rule out many of the applications we see.

Sampling error is mentioned and addressed in various places. As the focus of the manuscript and study is on survey data, it might make sense to refer and relate the work to the Total Survey Error (TSE) framework (see, e.g., Biemer, 2010).

We agree that this is a useful framework and that psychologists would benefit from using such an integrated framework rather than the separate focus on sample size and measurement error that characterises psychological study planning. However, we feel making this point well here is a bridge too far and may not properly fit into this manuscript (see also our response to your question about the Sampling Plan).

For a full paper, I would also suggest discussing how this work relates to broader discussions of how AI can influence and maybe also improve research in psychology and the social and behavioral sciences more generally (see, e.g., Bail, 2024; Demszky et al., 2023). In that, it would also be important to take into account and refer to more critical recent takes on the use of AI in science (e.g., Messeri & Crockett, 2024).

We agree with this and will speak about these points in the discussion when the data are in. Briefly, we would focus on the question of openness and transparency, the human-in-the-loop principle and various forms of the "stochastic parrot" critique. We see utility for AI, as we have pointed out, but as will also become clear, we do not think LLMs should replace data collection with real humans (compare e.g. https://syntheticusers.com/).

Finally, on a somewhat more political note: Can the use case of this study be (yet) a(nother) argument for not making psychological scales proprietary (esp. if their development is paid for by public funding)? The argument I see here is that we need open (source) models as well as open training data, and, in this case, open training data means open access scales.

> *We definitely agree with this argument and see our work as taking a notably different approach than other work (e.g., by Schoenegger et al., who use proprietary data and algorithms that they cannot share). But we also hope that averaged sentence vector representations can be a way to represent proprietary scales in search interfaces without violating intellectual property rights. Again, we would want to touch on this in the Discussion.*

Literature cited in this review

Bail, C. A. (2024). Can Generative AI improve social science? Proceedings of the National Academy of Sciences, 121(21). https://doi.org/10.1073/pnas.2314021121

Biemer, P. P. (2010). Total Survey Error: Design, Implementation, and Evaluation. Public Opinion Quarterly, 74(5), 817–848. https://doi.org/10.1093/poq/nfq058

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. Nature Reviews Psychology. https://doi.org/10.1038/s44159-023-00241-5

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. Nature, 627(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

Veselovsky, V., Ribeiro, M. H., Cozzolino, P., Gordon, A., Rothschild, D., & West, R. (2023). Prevalence and prevention of large language model use in crowd work. https://doi.org/10.48550/ARXIV.2310.15683

Replies to Hu Chuan-Peng, Reviewer 3

I read this manuscript with great interest.

The goal of this RR is to test the generalizability of the predictive power of a language model, "surveybot3000", by collecting a new dataset. The "hard" part of this study has finished and been tested with the pilot study. The sampling plan is clear and the strategy is straightforward.

> *We hope the reviewer is right and we are not in for a rude awakening.*

The only critical issue is the ambiguity in the hypotheses testing. In the design table, the authors described the following situations: the accuracy "matches or exceeds that found in the pilot study", "deteriorates but is still substantial", "halved", and "reduced below the accuracy of the pre-trained model", the words here are ambiguous. Using quantitative criteria will be more helpful. For example, employing an equivalence test to compare the correlations from the confirmative study to that of the pilot study.

> *We agree with the reviewer that more precise quantitative language is in order. We think equivalence tests are not needed because our point estimates are sufficiently precise that equivalence is unlikely, but we edited the Design Table to clarify the "buckets" into which correlations would fall and how we would discuss them.*

There are also a few minor issues:

(1) One of the osf links is invalid (https://osf.io/xbp8v), please re-check it.

> *We thank the reviewer for noticing this incorrect link. The corrected link is https://osf.io/bfhzy*

(2) The logic flow of the method section would be clearer if there was a roadmap for the whole study (model training → pilot study → confirmative study). This roadmap may help readers chunk the first few technical parts (pre-training and domain adaptation) together and focus more on the confirmative study, which is main body of the current RR.

> *We thank the reviewer for this great suggestion. We have adapted Figure 1 to include the pilot study and the planned confirmatory study.*

(3) Constrain the statement about "generalizability" in the future. Given the current training is largely based on English data, the utility of the model for other languages/samples needs further validation.

> *We absolutely agree with this point and have edited the texts to clarify these constraints on generalizability that we do not intend to break in this first study. We think the language constraint is important to mention. We do* ***expect*** *the model to generalize to new samples and different items, though of course we have yet to find out.*
>
> *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.*
>
> as well as
>
> 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 ~~data~~ English-language items.