

Peer-Review

Sankarlingam, Shivaani. 2026. "Self-Reported Mentalizing and AI Differentiation: An Empirical Study with Implications for Construct Validity." *Journal of High School Science* 10 (3): 56–84. <https://doi.org/10.64336/001c.164666>.

The manuscript addresses a timely and potentially important question: whether individual differences in mentalizing ability influence the ability to distinguish human-written from AI-generated text. The study is well written, the literature review is thorough, and the rationale is clearly articulated. However, I am unable to recommend acceptance in its current form because the data do not support the central hypothesis and the manuscript does not provide sufficient evidence for an alternative explanatory framework.

The primary hypothesis of the study is that mentalizing ability is associated with AI differentiation performance. However, neither aggregate good mentalizing (F1–F3) nor aggregate poor mentalizing (F4–F6) was significantly associated with overall AI differentiation, scientific abstract differentiation, or news headline differentiation. The manuscript therefore fails to support its principal research question. The only statistically significant finding reported is the negative association between reflexivity (F1) and news headline differentiation. Given the large number of correlations examined, this isolated result is vulnerable to multiple-comparison concerns. The manuscript does not report family-wise error correction or false-discovery-rate adjustment, making it unclear whether the finding would remain statistically significant after appropriate correction. Even if the F1 result remains significant, its effect size is small and explains only a limited fraction of the variance in AI differentiation performance. Consequently, the manuscript's discussion places considerable interpretive weight on a weak secondary result while the primary analyses are uniformly null.

A second concern is that the predictor variable is a self-report measure rather than a behavioral measure of mentalizing ability. The study therefore evaluates self-perceived mentalizing rather than demonstrated mentalizing performance. This distinction substantially limits the strength of any conclusions regarding the role of mentalizing in AI differentiation. A null relationship between self-reported mentalizing and AI differentiation cannot be interpreted as evidence that mentalizing itself is unrelated to AI differentiation.

A third concern is that the manuscript aggregates the six MMQ factors into broad "good" and "poor" mentalizing constructs without establishing that these aggregate scores are the most informative representation of the underlying psychological structure. It is possible that aggregation obscures meaningful relationships among specific dimensions of mentalizing. As presented, the study demonstrates only that the chosen summary measures are not predictive.

For these reasons, I cannot recommend publication in its current form. The manuscript does not provide convincing evidence that mentalizing predicts AI differentiation, nor does it identify a robust alternative cognitive mechanism.

RECOMMENDATION

However, I believe there remains a potentially valuable contribution hidden within the dataset. Rather than treating the MMQ solely as two aggregate constructs, I encourage the authors to perform exploratory analyses designed to identify whether a previously unrecognized cognitive balance or latent factor predicts AI differentiation ability.

Specifically, the authors may consider examining:

A Good-to-Poor Mentalizing Ratio

$$\text{Ratio} = (F1 + F2 + F3) / (F4 + F5 + F6)$$

This analysis would test whether relative mentalizing balance is more informative than absolute mentalizing strength.

A Mentalizing Balance Index

$$\text{Balance} = [(F1 + F2 + F3) - (F4 + F5 + F6)] / [(F1 + F2 + F3) + (F4 + F5 + F6)]$$

This formulation measures the extent to which adaptive mentalizing tendencies outweigh maladaptive tendencies and may better reflect overall mentalizing calibration.

Reflexivity-Normalized Indices

Given that F1 was the only factor displaying a significant relationship, the authors should investigate whether reflexivity relative to maladaptive factors predicts AI differentiation more strongly than reflexivity alone. Examples include:

$$F1 / (F4 + F5 + F6)$$

or

$$(F1 - F4) / (F1 + F4)$$

These analyses would test whether the balance between introspective tendencies and interpersonal dysfunction is more important than either factor independently.

Latent-Factor or Principal-Component Analysis

Rather than imposing a predefined good-versus-poor structure, the authors could use principal-component analysis, exploratory factor analysis, or other latent-variable approaches to identify emergent cognitive dimensions within the MMQ. Such dimensions may prove more predictive of AI differentiation than the original factor groupings.

Should any of these alternative representations demonstrate a robust relationship with AI differentiation and survive appropriate correction for multiple testing, the manuscript would shift from reporting a largely negative result to identifying a potentially novel cognitive predictor of AI discrimination ability. Such a contribution would be substantially more compelling and theoretically informative than the current analysis.

Until such evidence is provided, however, the manuscript remains primarily a null-result study whose principal conclusions are not sufficiently supported to justify publication.

Reviewer 1:

The manuscript addresses a timely and potentially important question: whether individual differences in mentalizing ability influence the ability to distinguish human-written from AI-generated text. The study is well written, the literature review is thorough, and the rationale is clearly articulated. However, I am unable to recommend acceptance in its current form because the data do not support the central hypothesis and the manuscript does not provide sufficient evidence for an alternative explanatory framework.

The primary hypothesis of the study is that mentalizing ability is associated with AI differentiation performance. However, neither aggregate good mentalizing (F1–F3) nor aggregate poor mentalizing (F4–F6) was significantly associated with overall AI differentiation, scientific abstract differentiation, or news headline differentiation. The manuscript therefore fails to support its principal research question. The only statistically significant finding reported is the negative association between reflexivity (F1) and news headline differentiation. Given the large number of correlations examined, this isolated result is vulnerable to multiple-comparison concerns. The manuscript does not report family-wise error correction or false-discovery-rate adjustment, making it unclear whether the finding would remain statistically significant after appropriate correction.

I agree that the primary hypothesis wasn't supported. Accordingly, I substantially revised the manuscript to present this finding transparently rather than emphasizing small significant results.

To do this I:

- Rewrote abstract to not inflate findings
- Rewrote the Results and Discussion to emphasize the consistently null findings across all primary analyses AND conducted and included the results of the Benjamini Hochberg FDR corrections
- Removed strong language implying support for the original hypothesis

- Reframed the study as an investigation into multiple theoretically and statistically motivated representations of self-reported Mentalizing

Locations where I did this in the new manuscript:

- Introduction (Section 1.4)
- Results (Sections 3.2.4, 3.3, Table 1)
- Discussion (first two paragraphs, last two paragraphs before 4.1)
- Abstract

Even if the F1 result remains significant, its effect size is small and explains only a limited fraction of the variance in AI differentiation performance. Consequently, the manuscript's discussion places considerable interpretive weight on a weak secondary result while the primary analyses are uniformly null.

I softened the weight placed on these claims throughout my paper and focused much less on Reflexivity in my Discussion and Abstract. The Discussion now explicitly emphasizes that these findings related to it shouldn't be interpreted as robust evidence for Reflexivity as a predictor of AI Differentiation because it failed FDR correction and was consistently not supported throughout multiple statistical models. Most of these changes can be found in Section 3.3.1 or throughout the bulk of the Discussion.

A second concern is that the predictor variable is a self-report measure rather than a behavioral measure of mentalizing ability. The study therefore evaluates self-perceived mentalizing rather than demonstrated mentalizing performance. This distinction substantially limits the strength of any conclusions regarding the role of mentalizing in AI differentiation. A null relationship between self-reported mentalizing and AI differentiation cannot be interpreted as evidence that mentalizing itself is unrelated to AI differentiation.

I went through and made sure to strictly define the difference between perceived and true Mentalizing in my paper. I've also explicitly mentioned in my Limitations (Section 4.2) that the null findings cannot be interpreted as evidence against Mentalizing itself, and have listed one of the future directions as looking at MMQ using a behavioral measure (which i myself don't have access or qualification to use).

A third concern is that the manuscript aggregates the six MMQ factors into broad "good" and "poor" mentalizing constructs without establishing that these aggregate scores are the most informative representation of the underlying psychological structure. It is possible that aggregation obscures meaningful relationships among specific dimensions of mentalizing. As presented, the study demonstrates only that the chosen summary measures are not predictive.

This suggestion motivated the largest change to the manuscript: now, in order to evaluate whether aggregation obscured meaningful relationships, I performed several new exploratory analyses using my prior data:

- OLS multiple linear regression using all six MMQ factors simultaneously
- PCA to identify latent dimensions
- Bootstrap sampling
- Five-fold cross-validation
- Comparison of 3 competing regression models using AIC and BIC

These analyses allowed me to investigate whether alternative representations of Mentalizing predicted AI Differentiation more effectively than the original Good/Poor Mentalizing scores did.

Although none demonstrated significant performance, this convergence across multiple analytical frameworks substantially strengthens confidence that the primary findings are not simply artifacts of the original MMQ framework. I revised the entire paper to reflect this new framing of my results and intent.

For these reasons, I cannot recommend publication in its current form. The manuscript does not provide convincing evidence that mentalizing predicts AI differentiation, nor does it identify a robust alternative cognitive mechanism.

I have now included in section 4.1 recommendations based on previous literature and my own findings for new, promising directions for research.

RECOMMENDATION

However, I believe there remains a potentially valuable contribution hidden within the dataset. Rather than treating the MMQ solely as two aggregate constructs, I encourage the authors to perform exploratory analyses designed to identify whether a previously unrecognized cognitive balance or latent factor predicts AI differentiation ability.

Specifically, the authors may consider examining:

A Good-to-Poor Mentalizing Ratio $\text{Ratio} = (F1 + F2 + F3) / (F4 + F5 + F6)$

This analysis would test whether relative mentalizing balance is more informative than absolute mentalizing strength.

A Mentalizing Balance Index

$\text{Balance} = [(F1 + F2 + F3) - (F4 + F5 + F6)] / [(F1 + F2 + F3) + (F4 + F5 + F6)]$

This formulation measures the extent to which adaptive mentalizing tendencies outweigh maladaptive tendencies and may better reflect overall mentalizing calibration.

Reflexivity-Normalized Indices

Given that F1 was the only factor displaying a significant relationship, the authors should investigate whether reflexivity relative to maladaptive factors predicts AI differentiation more strongly than reflexivity alone. Examples include:

$F1 / (F4 + F5 + F6)$ or

$(F1 - F4) / (F1 + F4)$

These analyses would test whether the balance between introspective tendencies and interpersonal dysfunction is more important than either factor independently.

I conducted all of these procedures and found that the results were insignificant, and I couldn't find a suitable way to place them into my new manuscript without confusing the narrative, so I put them aside and focused on my other analyses.

Latent-Factor or Principal-Component Analysis

Rather than imposing a predefined good-versus-poor structure, the authors could use principal-component analysis, exploratory factor analysis, or other latent-variable approaches to identify emergent cognitive dimensions within the MMQ. Such dimensions may prove more predictive of AI differentiation than the original factor groupings.

I implemented this recommendation by conducting a Principal Component Analysis of the six MMQ factors, and although PC2 demonstrated an initial association with AI Abstract Differentiation, this correlation didn't survive FDR correction and therefore was interpreted cautiously. I created a new section in my Methods and Results sections (2.3.3 and 3.3.2) to explain and discuss my PCA (alongside the other analyses) in-depth, as well as more explanation in the Discussion.

Should any of these alternative representations demonstrate a robust relationship with AI differentiation and survive appropriate correction for multiple testing, the manuscript would shift from reporting a largely negative result to identifying a potentially novel cognitive predictor of AI discrimination ability. Such a contribution would be substantially more compelling and theoretically informative than the current analysis.

Although none of the alternative models showed a robust relationship with AI Differentiation after correction and model validation, I believe these analyses nevertheless strengthened the manuscript by demonstrating that the predominantly null findings remain consistent across multiple theoretically and statistically motivated representations of self-reported Mentalizing.

To highlight this, the manuscript was reframed to emphasize methodological robustness and the importance of evaluating psychological hypotheses using multiple complementary analytical models, rather than relying on a single statistical representation. Until such evidence is provided, however, the manuscript remains primarily a null-result study whose principal conclusions are not sufficiently supported to justify publication.

Reviewer 2:

Please verify that all links to the references point to the correct source.

All links to references have been checked and I updated some links and citations to be style-specific and correct.

The authors must disclose and acknowledge any assistance received in the preparation of this manuscript, including but not limited to editorial, technical, analytical, or writing support. All such contributions must be clearly stated in the Acknowledgments section.

An acknowledgements section was created and placed after the end of the Discussion, before the References.

Include enough recent references along with foundational ones. Ensure references directly support your claims. Avoid "padding." Use credible sources (peer-reviewed journals, reputable books, official reports).

All sources are peer-reviewed or come from official textbooks. The padding issue was addressed throughout the manuscript, as mentioned with the previous reviewer.

All assumptions (including implicit assumptions) must be explicitly stated and clearly justified. The authors should explain why each assumption is reasonable and discuss its impact on the results and conclusions.

More explanation was included in the Literature Review (1.4) and Methods (2.3.2, 2.3.3) to justify the new directions and exploratory analyses.

Avoid overreaching conclusions that extend beyond what is supported by the data and analysis.

Most of the Discussion has been edited to avoid overarching claims and definitive language.

Verify that you have used past perfect tense and third person throughout the manuscript wherever applicable.

I edited the manuscript for tense and caught a few present tense slips which have been corrected.

Thank you for addressing my comments so comprehensively. I have an additional comment about concept and construct. See below and please address in the manuscript.

1. I would like you to include this content (about your construct itself) in your limitations section and also briefly discuss in the discussion and conclusion sections

“A more fundamental limitation of the present work concerns the construct validity of “AI differentiation ability” itself. Unlike many psychological tasks that measure responses to relatively stable human behaviors or cognitive phenomena, AI differentiation involves identifying outputs generated by rapidly evolving algorithmic systems. As large language models continue to change in architecture, training data, and stylistic characteristics, the cognitive processes required for successful differentiation may likewise change over time. Consequently, it remains uncertain whether AI differentiation represents a stable human cognitive construct with enduring psychological predictors or whether it reflects performance against a temporally specific technological artifact. The present findings should therefore be interpreted within the context of the specific LLM generation and experimental paradigm employed, and future work should further establish the construct validity and temporal stability of AI differentiation before drawing broader conclusions about its cognitive determinants.”

Also,

"The manuscript implicitly treats AI-generated text as though it represents a natural category analogous to human behavior. However, AI-generated language is an engineered artifact whose observable characteristics reflect transient design choices made by developers rather than enduring properties of intelligence itself. As successive generations of LLMs are explicitly optimized to eliminate detectable stylistic artifacts, the cognitive demands of distinguishing AI from human writing are expected to evolve correspondingly. This raises a fundamental question regarding whether “AI differentiation ability” constitutes a stable psychological construct suitable for identifying enduring cognitive predictors, or whether it instead reflects performance against a continually changing technological target. The manuscript should explicitly acknowledge this conceptual limitation.

Therefore, you should also soften statements like: “These findings narrow the search for cognitive affectors of AI attribution.”

Because you haven’t necessarily narrowed the search. You would have narrowed it only if AI differentiation were to be a valid, stable construct. If the construct itself is unstable, then today’s negative result may have little bearing on tomorrow’s LLMs. A more appropriate conclusion would be that you found no robust association under the specific technological conditions studied. That is a more defensible interpretation.

2. The bootstrap analysis reports that Good Mentalizing has a confidence interval excluding zero for the News Headline model, suggesting stability, while the five-fold cross-validation shows negative out-of-sample R^2 , indicating poor predictive performance. These results are not inherently inconsistent, but the manuscript should explicitly explain that bootstrap assesses the stability of parameter estimates within the sampled population, whereas cross-validation evaluates predictive

generalizability to new data. Without that explanation, readers may perceive an apparent contradiction.

Summary:

Conceptual limitation regarding construct validity. While the revised manuscript appropriately tempers its conclusions regarding self-reported mentalizing, I believe a more fundamental limitation remains insufficiently acknowledged. The study implicitly treats “AI differentiation ability” as a stable psychological construct for which enduring cognitive predictors can be identified. However, unlike most psychological phenomena, the target being differentiated is not a stable natural phenomenon but the output of a rapidly evolving algorithmic system. As successive generations of large language models are intentionally redesigned to eliminate recognizable stylistic artifacts, the cognitive processes required to distinguish AI from human writing may also change. Consequently, it remains uncertain whether AI differentiation represents a stable cognitive ability or performance against a temporally specific technological artifact. The present findings should therefore be interpreted as applying to the particular LLM generation, prompting strategy, and experimental paradigm used in this study. I encourage the authors to explicitly discuss this issue in the Limitations and moderate any statements suggesting that the present findings meaningfully narrow the search for enduring cognitive predictors of AI differentiation.

Thank you for addressing my comments so comprehensively. I have an additional comment about concept and construct. See below and please address in the manuscript.

I would like you to include this content (about your construct itself) in your limitations section and also briefly discuss in the discussion and conclusion sections
“A more fundamental limitation of the present work concerns the construct validity of “AI differentiation ability” itself. Unlike many psychological tasks that measure responses to relatively stable human behaviors or cognitive phenomena, AI differentiation involves identifying outputs generated by rapidly evolving algorithmic systems. As large language models continue to change in architecture, training data, and stylistic characteristics, the cognitive processes required for successful differentiation may likewise change over time. Consequently, it remains uncertain whether AI differentiation represents a stable human cognitive construct with enduring psychological predictors or whether it reflects performance against a temporally specific technological artifact. The present findings should therefore be interpreted within the context of the specific LLM generation and experimental paradigm employed, and future work should further establish the construct validity and temporal stability of AI differentiation before drawing broader conclusions about its cognitive determinants.”

Also,

"The manuscript implicitly treats AI-generated text as though it represents a natural category analogous to human behavior. However, AI-generated language is an engineered artifact whose observable characteristics reflect transient design choices made by developers rather than enduring properties of intelligence itself. As successive generations of LLMs are explicitly optimized to eliminate detectable stylistic artifacts, the cognitive demands of distinguishing AI from human writing are expected to evolve correspondingly. This raises a fundamental question regarding whether “AI differentiation ability” constitutes a stable psychological construct suitable for identifying enduring cognitive predictors, or whether it instead reflects performance against a continually changing technological target. "The manuscript should explicitly acknowledge this conceptual limitation. Therefore, you should also soften statements like: “These findings narrow the search for cognitive affectors of AI attribution.”

Because you haven’t necessarily narrowed the search. You would have narrowed it only if AI differentiation were to be a valid, stable construct. If the construct itself is unstable, then today’s negative result may have little bearing on tomorrow’s LLMs. A more appropriate conclusion would be that you found no robust association under the specific technological conditions studied. That is a more defensible interpretation.

I’ve combined important parts of both suggested paragraphs and included them in section 4.2, as well as editing the beginning of the Discussion to match. The aforementioned sentence in the Conclusion was accordingly softened, and a mention of the limitations of using AI as a construct was also added. I have also added a similar phrase to the Abstract to match. The bootstrap analysis reports that Good Mentalizing has a confidence interval excluding zero for the News Headline model, suggesting stability, while the five-fold cross-validation shows negative out-of-sample R^2 , indicating poor predictive performance. These results are not inherently inconsistent, but the manuscript should explicitly explain that bootstrap assesses the stability of parameter estimates within the sampled population, whereas cross-validation evaluates predictive generalizability to new data. Without that explanation, readers may perceive an apparent contradiction.

A new paragraph 4 was added to the Discussion to make this clarification.

Thank you for addressing my comments.

Before I recommend acceptance, I would like you to

1. submit a .docx file of the manuscript, formatted correctly to JHSS specifications (see website).

2. Change the title to better reflect your final product. Something along the lines of “Self-Reported Mentalizing and AI Differentiation: An Empirical Study with Implications for Construct Validity” or similar.

3. You did not explicitly address my F1-F6 subset comment. This is OK since you have amply addressed this with alternative methodology. However, I would like to see this as a limitation. Something along the lines of “Although alternative statistical representations of the MMQ were explored, the manuscript does not investigate whether alternative psychologically motivated combinations of the six MMQ dimensions might outperform the predefined Good/Poor composites. Such analyses could be an interesting direction for future work but are not essential for the conclusions of the present study.”

4. Moderate this sentence in the ‘limitations’ section “Though the present study considers AI Differentiation ability as a natural category similar to an innate human behavior...” to “The present study necessarily operationalizes AI Differentiation as a measurable human ability for the purposes of empirical investigation; however, whether this represents a stable psychological construct remains uncertain.”

5. Please provide better resolution on your images. They appear blurry at present.

Thank you for addressing my comments.

Before I recommend acceptance, I would like you to

submit a .docx file of the manuscript, formatted correctly to JHSS specifications (see website).

Formatting of headings/subheadings has been changed, manuscript has been downloaded and submitted as a .docx file.

Change the title to better reflect your final product. Something along the lines of “Self-Reported Mentalizing and AI Differentiation: An Empirical Study with Implications for Construct Validity” or similar.

Title has been changed accordingly.

You did not explicitly address my F1-F6 subset comment. This is OK since you have amply addressed this with alternative methodology. However, I would like to see this as a limitation. Something along the lines of “Although alternative statistical representations of the MMQ were explored, the manuscript does not investigate whether alternative psychologically motivated combinations of the six MMQ dimensions might outperform the predefined Good/Poor composites. Such analyses could be an interesting direction for future work but are not essential for the conclusions of the present study.”

A derivative of the aforementioned sentence was added to the final paragraph of section 4.2.

Moderate this sentence in the ‘limitations’ section “Though the present study considers AI Differentiation ability as a natural category similar to an innate human behavior...” to “The present study necessarily operationalizes AI Differentiation as a measurable human ability for the purposes of empirical investigation; however, whether this represents a stable psychological construct remains uncertain.”

The sentences referred to have been softened accordingly in the third paragraph of section 4.2.

Please provide better resolution on your images. They appear blurry at present.

I recompiled the images as .pngs so that they're hopefully higher resolution.

Thank you for addressing my comments. Accepted.