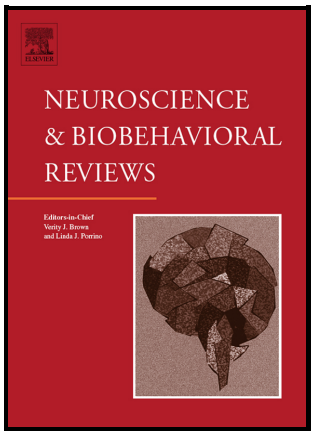
Three threats to the validity of the Reading the Mind in the Eyes test: A commentary on Pavlova and Sokolov (2022)

Wendy C. Higgins, Robert M. Ross, Vince Polito, David M. Kaplan



PII: S0149-7634(23)00057-X

DOI: https://doi.org/10.1016/j.neubiorev.2023.105088

Reference: NBR105088

To appear in:
Revised date: 8 February 2023
Accepted date: 9

Please cite this article as: Wendy C. Higgins, Robert M. Ross, Vince Polito and David M. Kaplan, Three threats to the validity of the Reading the Mind in the Eyes test: A commentary on Pavlova and Sokolov (2022), , (2022) doi:https://doi.org/10.1016/j.neubiorev.2023.105088

Three threats to the validity of the Reading the Mind in the Eyes test: A commentary on Pavlova and Sokolov (2022).

Wendy C. Higgins[a*1], Robert M. Ross[b], Vince Polito[a], David M. Kaplan[a]

[a] Macquarie University, School of Psychological Sciences, NSW 2109, Australia

[b] Macquarie University, Department of Philosophy, NSW 2109, Australia

Author Note

Wendy C. Higgins https://orcid.org/0000-0003-1357-8330
wendy.higgins@mq.edu.au

David M. Kaplan https://orcid.org/0000-0001-7532-1114
david.kaplan@mq.edu.au

Vince Polito https://orcid.org/0000-0003-3242-9074
vince.politio@mq.edu.au

Robert M. Ross https://orcid.org/0000-0001-8711-1675
robert.ross@mq.edu.au

When scientists lack validity evidence for measures, they lack the necessary information to evaluate the overall validity of a study's conclusions.

~ Flake and Fried (2020, p. 457)

In their ambitious review of the Reading the Mind in the Eyes Test (RMET) literature, Pavlova and Sokolov (2022) draw attention to many important issues including a critical question: "what does the RMET truly measure?" (p. 3). Unfortunately, when addressing this question, they provide limited discussion of the structural component of the RMET's construct validity, despite it being vital to determining what this test truly

[1] Correspondence concerning this article should be addressed to Wendy C. Higgins, Macquarie University, NSW, 2019, Australia. Email: wendy.higgins@mq.edu.au

measures. This is unfortunate given the growing evidence of a "measurement crisis" in psychology, according to which researchers routinely fail to rigorously examine whether psychological measures truly measure what they aim to measure (Flake & Fried, 2020). In this commentary, we briefly summarise research that suggests that the RMET fails to exhibit acceptable levels of structural validity, which brings much of the RMET literature into question. In particular, we highlight three key limitations that we recently identified (Higgins et al., 2022).

First, the RMET does not have a consistent factor structure. A noteworthy limitation of Baron-Cohen et al.'s (2001) presentation of the RMET is that it does not include a factor analysis. Subsequent tests of the dimensionality of the RMET are rare, and those few factor analyses that have been conducted either found poor model fit or failed to report key statistics for assessing model fit (Higgins et al., 2022). Nonetheless, the RMET is routinely interpreted as measuring a single underlying psychological construct. To rigorously examine the factor structure of the RMET, we conducted what we believe is the first pre-registered psychometric study focused on the structural validity of the English language RMET (Higgins et al., 2022). In a demographically representative study of 1,181 Americans aged 18 to 88, we failed to identify a well-fitting factor model. Moreover, all models that we tested had weak factor loadings, multiple items failing to load on to any factor, and we could not identify any conceptual explanation for the multiple factors.

Second, the RMET often demonstrates poor internal consistency. A recent meta-analysis based on the most frequently reported statistic of internal consistency, Cronbach's alpha, found that half of reported values were below the conventional minimum acceptable value of .70 (Kittel et al. 2021). While the meta-analytic point estimate was an "acceptable" .73, the 95% confidence interval included unacceptably low alpha values [.65, .79]. Moreover, because the RMET is long, and studies have found weak – and even negative – interitem correlations, Kittel et al. suggested that the acceptable point estimate

is likely an artifact of test length (i.e., adding items to a test typically increases alpha and with 36 items, the RMET is a long test).

Third, there is considerable uncertainty about what mental states are being expressed in the photos used for each RMET item and whether each item has a "correct" response. This is because Baron-Cohen et al. (2001) selected photos for the RMET from magazines without knowledge of what emotions the models were feeling or attempting to express, then chose four mental state terms – the "correct" response and three "incorrect" foils – for each candidate RMET item and validated them through consensus in a sample of 103 University of Cambridge undergraduate students and 122 members of the general public in the UK. They used two validation criteria that must be passed for a candidate item to be retained: at least 50% of participants must select the "correct" target and no more than 25% of participants select the same "incorrect" foil. Importantly, if this consensus does not hold across samples, then it is questionable whether the "correct" responses generalise to different populations. Unfortunately, individual item response rates are rarely reported in RMET studies, making it impossible to estimate how often consensus levels are maintained across the diverse populations sampled in RMET research. Troublingly, our psychometric study (Higgins et al., 2022) found that eight of the 36 items (i.e., 22%) failed to meet at least one of the two original validation criteria.

To their credit, Pavlova and Sokolov (2022) do briefly mention some concerns associated with the structural validity of the RMET. In particular, they note that internal consistency can be low and state that "There are also several points regarding the psychometric characteristics of the RMET . . . and multidimensionality of the response outcome" (p. 3), citing several papers. However, they avoid the elephant in the room: the uncertain factor structure, poor internal consistency, and concerns about whether items have "correct" responses make it far from clear that the RMET measures an ability to read emotions from the eyes. This lack of careful engagement with validity concerns manifests

in several places. For example, Pavlova and Sokolov (2022) summarised a study that translated two versions of the RMET into Korean, one that retained the original items and one that used Asian eyes (Koo et al., 2021), observing that these versions were "comparable in terms of psychometric properties" (Pavlova & Sokolov, 2022, p. 8). However, they did not report that these comparable Cronbach's alpha values were both .54, which is well below what is considered acceptable for a valid psychological test. This downplaying of evidence against the validity of the RMET is consistent with the "measurement schmeasurement" attitude that Flake and Fried (2020) identify across the psychology literature, with uncertainty about the validity of the RMET creating uncertainty about the conclusions of studies throughout the vast literature that uses it. Going forward, we argue that it is *crucial* that scholars carefully examine the validity of the RMET to come to a clearer understanding of what knowledge and skills participants are tapping into when they select the "correct" mental state descriptors for these pairs of eyes.

Acknowledgments

WCH is funded by an Australian Government Research Training Program (RTP) Scholarship and a Macquarie University Research Excellence Scholarship, RMR is funded by a Macquarie University Research Fellowship and the John Templeton Foundation (Grant ID: 62631), and VP is funded by a Macquarie University Research Fellowship.

References

References:

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *42*(2), 241–251. http://dx.doi.org/10.1111/1469-7610.00715

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

Higgins, W. C., Ross, R. M., Langdon, R., & Polito, V. (2022). The "Reading the Mind in the Eyes" Test shows poor psychometric properties in a large, demographically representative U.S. sample. *Assessment*, *0*(0). https://doi.org/10.1177/10731911221124342

Kittel A. F. D., Olderbak S., Wilhelm O. (2021). Sty in the Mind's Eye: A meta-analytic investigation of the nomological network and internal consistency of the "Reading the Mind in the Eyes" test. *Assessment*, 29(5), 872–895. https://doi.org/10.1177/1073191121996469

Koo, S.J., Kim, Y.J., Han, J.H., Seo, E., Park, H.Y., Bang, M., Park, J.Y., Lee, E., An, S.K. (2021). "Reading' the Mind in the Eyes Test": Translated and Korean versions. *Psychiatry Investigation*, *18*(4), 295–303. https://doi.org/10.30773/pi.2020.0289.

Pavlova, & Sokolov, A. A. (2022). Reading language of the eyes. *Neuroscience and Biobehavioral Reviews*, *140*, 104755–104755. https://doi.org/10.1016/j.neubiorev.2022.104755