Assessing Bias in Original & Updated Reading the Mind in the Eyes Test (RMET)

Abstract: The Reading the Mind in the Eyes Test (RMET) is a prevalent clinical measure for social cognition and theory of mind. However, due to the exclusive use of monochromatic pictures of white individuals and unintuitive vocabulary in the questionnaire, the original RMET has been updated to include diverse, full-color photographs of male and female faces with simpler vocabulary. In this study, we assess whether these revisions meaningfully address identified areas of potential bias in the original RMET. Using model selection, linear regression, measures of model fit, ANOVA, and post-hoc Tukey HSD, we found that ethnicity, gender, and native language were equally predictive in both the original and updated versions of the RMET, with non-European individuals consistently scoring the lowest. Our findings suggest that inherent qualities of the RMET, aside from choice of vocabulary and ethnicities of the photographed faces, may contribute to the biased predictivity of demographics on RMET performance.

1. Background & Significance

In research and clinical practice, the Reading the Mind in the Eyes Test (RMET) has been an important neuropsychiatric measure for testing theory of mind, social cognition, and autism (Voracek & Dressler, 2006; Richell et al., 2003). In this test, participants are presented with photographs of the human eye region and are asked to match the mental state of the person with one of four possible mental state words. However, the RMET may demonstrate critical psychometric flaws. First, the measure only employs black-and-white pictures of white persons to test cognitive empathy and theory of mind across a diverse population (Appendix 6.1). Second, there was a greater representation of female faces—many of which were unnaturally depicted (i.e., makeup)—compared to male faces. Third, it employs unintuitive sets of vocabulary that are esoteric in everyday language. Given these critical limitations, the RMET was revised to include images of natural male and female faces across a variety of races and ethnicities. Additionally, mental state terms have been changed to be more intuitive. But does this updated RMET really eliminate the demographic biases identified in the original RMET?

2. Hypothesis

We hypothesize that the updated RMET will show less signs of demographic bias compared to the original RMET by demonstrating that certain demographic variables—such as ethnicity, gender, native language, and education—do not predict the updated RMET scores as well as the original RMET scores. We further hypothesize that non-European groups will fare better in the new version of the RMET compared to the older version of the RMET.

3. Methods

3.1 Data Collection

The data were obtained from the Laboratory for Brain and Cognitive Health Technology, part of the Institute for Technology in Psychiatry at McLean Hospital and Harvard Medical School. The subjects were 4820 test takers on TestMyBrain.org, an internet-based research platform that allows users to participate in cognitive tests for free. The dataset includes demographic information and scores on the original and updated versions of the RMET (Olderbak et al., 2015). Each participant completed a mixed version of the RMET that included items from both the original and updated tests. Two equivalent versions of the aforementioned test were administered to two separate samples, one of which was later used as a training dataset and one of which was later used as the testing dataset.

3.2 Variable Creation

Scores for the original and updated tests were computed and disaggregated by calculating percent correct of the participants' responses to items from each test. For ethnicities, we were interested in comparing differences between white and non-white individuals to explore bias against people of color. Because we wanted to ensure a large enough sample size in each group analyzed, ethnicities were grouped into "European," "Non-European," and "Mixed" categories from multiple selections including "Africa," "Americas," "Asian," "Europe," "Pacific," "Uncertain," or "Decline." A "Missing" category included users who selected "Uncertain" or "Decline." For education, we identified missingness and created a category for these users.

3.3 Data Analysis

First, in order to assess the reliability of the original and updated versions of the RMET, internal reliability was calculated for each test using split-half correlation and a Spearman-Brown correction. Then, in order to compare the relationship of performance between the two versions

of RMET, a correlation of scores was computed. Finally, a paired t-test comparing the mean scores of all subjects on the original test and the mean scores of all subjects on the updated test was performed.

Next, in order to systematically choose predictive demographic characteristics for statistical analysis, we fit four linear regression models predicting the scores of the original RMET based on combinations of gender, ethnicity, native language, and education in the training dataset. We compared these models and selected for predictors that led to a model with the lowest AIC, BIC, and residual variance for use in further analysis. The chosen model included linear terms for ethnicity, gender, and whether native language was English (Appendix 6.2).

We evaluated model fit on a second set of participants' original and updated RMET scores by calculating the root mean squared errors (RMSE). An ANOVA was run to determine whether any predictors significantly explained the variance in scores. If predictors significantly explained variance in scores, pairwise comparisons were explored through a post-hoc Tukey HSD test. Assumptions for all performed statistical tests (e.g. independence, normality, lack of outliers, linearity) were met.

4. Results

Computation of split-half internal reliability (Cronbach's alpha) of the original and updated tests revealed that the original test maintained a lower internal reliability (α = .51) compared to the updated test (α = .60). Each participant's original and updated scores were found to be moderately correlated (r = 0.4), which replicated in a second set of subjects (r = 0.51) (Figure 2).

A paired t-test revealed that the overall mean for original scores (M = 0.664, SD = 0.160) is significantly greater than the overall mean for updated scores (M = 0.623, SD = 0.168, p < .0001); with 95% confidence, the true mean difference in scores between original and updated RMET scores within each test taker is captured by the interval (-0.046, -0.036).

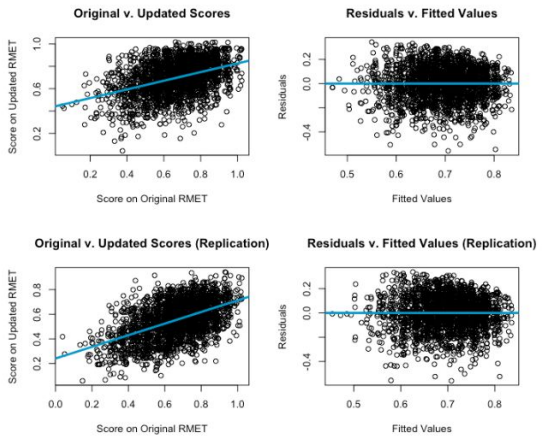


Figure 2: Correlation of scores on the original and updated RMET. There appears to be a moderately linear relationship between the original and updated scores of RMET, consistent with the intended similarity of the two versions of the RMET.

We trained a linear regression model of scores predicted by ethnicity, gender, and native language on the original RMET scores of the training dataset. The RMSE of 0.154 of the model on the original scores (testing data) was very similar to the resulting RMSE of 0.153 in the initially trained data, supporting that there is no evidence for overfitting, and that the trained model fit equally well to both the train and test datasets of the original RMET scores. On the contrary, the RMSE of these participants' updated RMET scores was larger (0.179). A larger RMSE suggests that the relationship between scores and chosen demographic predictors on the updated RMET was different than that of the original RMET, leading to greater degree of misprediction of updated RMET scores based on the trained linear regression model.

To understand the exact manner by which the relationship between predictors and scores have changed in the updated RMET, an analysis of variance (ANOVA) was performed. Ethnicity, gender, and native language significantly explained the variance in both the original and updated versions (p < .0001). A post-hoc Tukey test showed that the non-European group

differed significantly from the European and mixed-European groups ($p < .0001$). Additionally, male subjects differ significantly from both female and genderqueer subjects on both versions ($p < .01$). Finally, native English speakers differed significantly from non-native English speakers on both versions (Figure 3, Appendix 6.3). Surprisingly, we observed that the non-European group scored the lowest of all groups in all versions (both the original and revised) of RMET, consistent with other findings (Prevost et al., 2013).

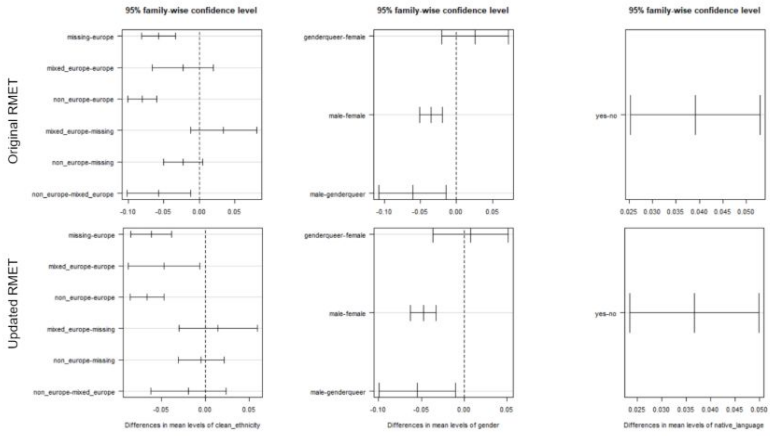


Figure 3: Results from pairwise Tukey HSD tests. Line and bar represents mean and 95% CI of the differences between indicated groups.

5. Discussion

First, on a gross level, we observed that the updated RMET led to lower overall scores in most demographic groups. However, we observed that the non-European demographic group scored most poorly among all ethnic groups on the revised RMET, just as was observed in the original RMET scores (Figure 3, Appendix 6.3). The lower scores of non-European individuals, we conclude, is not necessarily caused by the presence of homogeneously white face as stimuli; rather, the lower scores should be attributed to another inherent quality of the test. For example, we argue that the typical, citizen-science consensus method for determining the "correct" answer for the test—entailing the recruitment of a majority of ethnically white and European participants—may have contributed to a systematic bias against ethnic minorities on their performance in the RMET (Bjornsdottir et al., 2016). Because "correct" answers may be based on a eurocentric measure of emotional expression, alternate scoring methods should be examined (Adams et al., 2010). In addition, our finding that a linear regression model trained on the original RMET with ethnicity, gender, and native language terms was less predictive of scores on updated RMET suggests that the relationship between the aforementioned demographic predictors and RMET score has changed in the new RMET version in some way, but not in the anticipated manner (i.e., reducing demographic bias). Regardless, the updated RMET appears to maintain improved psychometric properties comparative to the old RMET: the updated RMET appears to maintain higher internal reliability, though still only moderate, and the scores on the original and new tests correlate moderately (Fernández-Abascal et al., 2013).

5.1 Limitations & Future Directions

The sample population was self-selected by use of an online testing platform and, it is difficult to know whether these results can be generalized to the broader population, especially given that the vast majority of individuals in the dataset had received at least some college education, unlike the general public. In addition, although questions in the two replicated tests were of similar format and caliber, subjects who took one version of the combined original and updated test (train dataset) tended to perform worse on the updated section, whereas the other cohort (test dataset) performed worse on the original section. Further analysis regarding these differences, in addition to examining the eurocentric effect in scoring methods, are warranted.

References

Adams, R. B., Rule, N. O., Franklin, R. G., Wang, E., Stevenson, M. T., Yoshikawa, S., … Ambady, N. (2010). Cross-cultural Reading the Mind in the Eyes: An fMRI Investigation. *Journal of Cognitive Neuroscience*, *22*(1), 97–108. doi: 10.1162/jocn.2009.21187

Bjornsdottir, R. T., & Rule, N. O. (2016). On the relationship between acculturation and intercultural understanding: Insight from the Reading the Mind in the Eyes test. *International Journal of Intercultural Relations*, *52*, 39–48. doi: 10.1016/j.ijintrel.2016.03.003

Fernández-Abascal, E. G., Cabello, R., Fernández-Berrocal, P., & Baron-Cohen, S. (2013). Test-retest reliability of the 'Reading the Mind in the Eyes' test: A one-year follow-up study. *Molecular Autism*, *4*(1), 33. doi: 10.1186/2040-2392-4-33

Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brenneman, M. W., & Roberts, R. D. (2015). A psychometric analysis of the reading the mind in the eyes test: Toward a brief form for research and applied settings. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.01503

Prevost, M., Carrier, M.-E., Chowne, G., Zelkowitz, P., Joseph, L., & Gold, I. (2013). The Reading the Mind in the Eyes test: Validation of a French version and exploration of cultural variations in a multi-ethnic city. *Cognitive Neuropsychiatry*, *19*(3), 189–204. doi: 10.1080/13546805.2013.823859

Richell, R., Mitchell, D., Newman, C., Leonard, A., Baron-Cohen, S., & Blair, R. (2003). Theory of mind and psychopathy: Can psychopathic individuals read the 'language of the eyes'? *Neuropsychologia*, *41*(5), 523–526. doi: 10.1016/s0028-3932(02)00175-6

Voracek, M., & Dressler, S. G. (2006). Lack of correlation between digit ratio (2D:4D) and Baron-Cohen's "Reading the Mind in the Eyes" test, empathy, systemising, and autism-spectrum quotients in a general population sample. *Personality and Individual Differences, 41*(8), 1481–1491. doi: 10.1016/j.paid.2006.06.009

6. *Appendix*

6.1: Examples of stimuli from the Reading the Mind in the Eyes Test: The original test (left) used black-and-white pictures of white subjects taken from magazines. The updated test (right) uses full-color images of multiracial subjects, intended to be more similar to those encountered in real life.

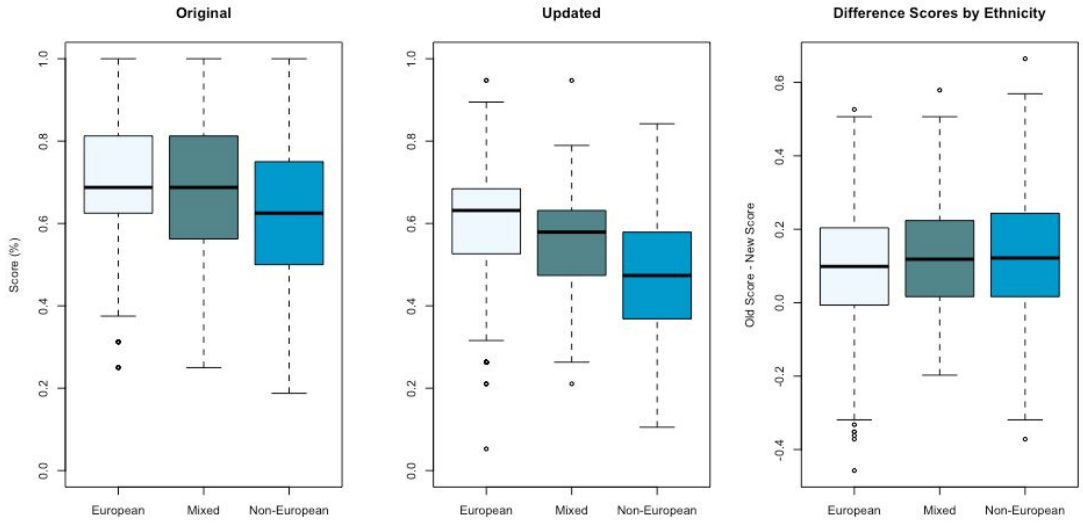
6.2 : Models compared in model selection. Criteria for comparisons between predictive models are included in the table, and the chosen model is highlighted.

Model Selection					
	Residual variance	**R^2**	**Adjusted R^2**	**BIC**	**AIC**
Ethnicity	0.02426	0.04731	0.04609	−2020.444	−2049.225
Ethnicity and Gender	0.02390	0.06156	0.05954	−2040.128	−2080.421
Ethnicity, Gender, and Native Language	0.02356	0.07471	0.07233	−2065.344	−2111.394
Ethnicity, Gender, Native Language, and Education	0.02269	0.09588	0.09012	−1976.903	−2068.144

6.3: Boxplot and Tukey HSD results of original and updated RMET scores by (A) ethnicity, (B) gender, and (C) native language.

(A) Ethnicity Subgroups

i. Boxplots



Ii. P-values of post-hoc Tukey HSD

Original RMET

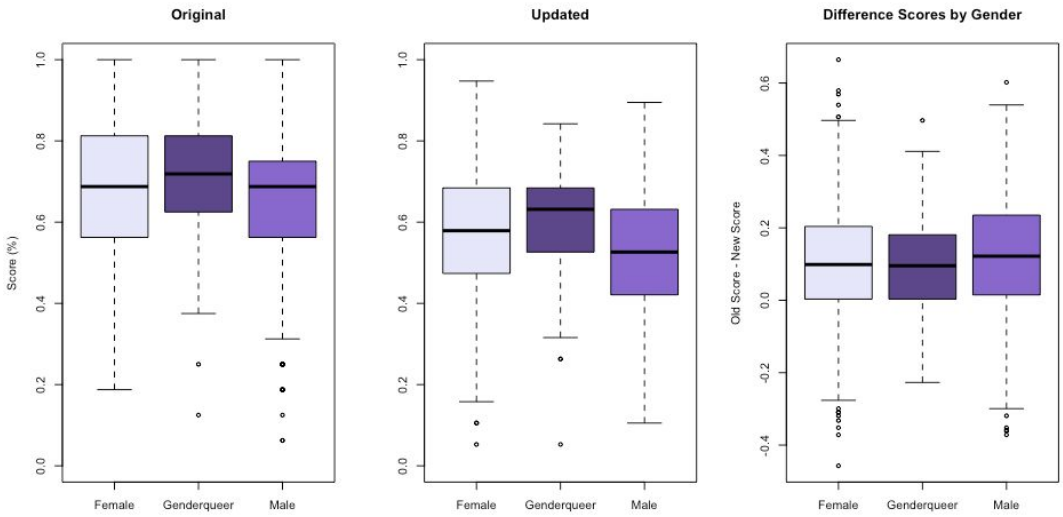
	European	Mixed	Non-European
European		$p = .98$	$p < .0001$
Mixed			$p < .0001$
Non-European			

Updated RMET

	European	Mixed	Non-European
European		$p = .13$	$p < .0001$
Mixed			$p < .0001$
Non-European			

(B) Gender Subgroups

i. Boxplots



Ii. P-values of post-hoc Tukey HSD

Original RMET

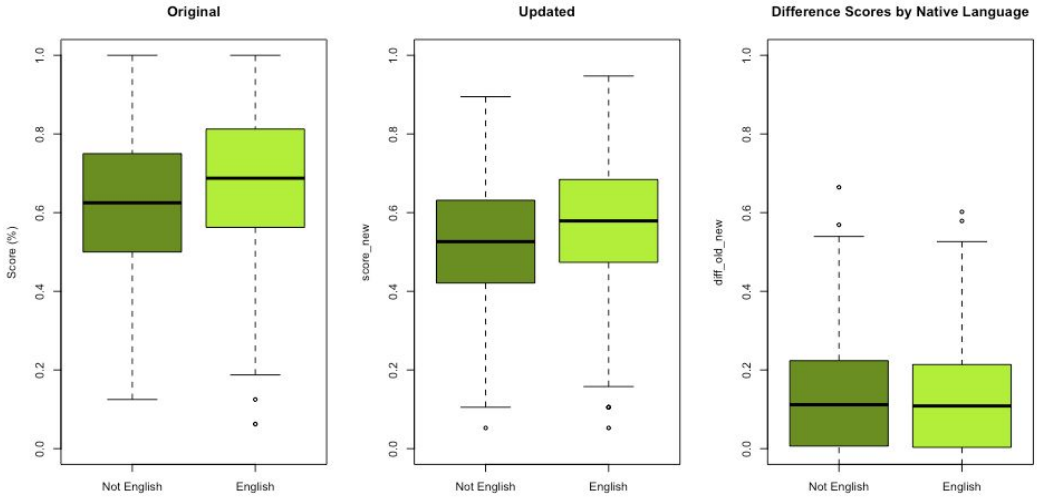
	Female	Gender-queer	Male
Female		$p = .23$	$p < .0001$
Gender-queer			$p = .002$
Male			

Updated RMET

	Female	Gender-queer	Male
Female		$p = .18$	$p < .0001$
Gender-queer			$p < .0001$
Male			

(C) Language Subgroups

i. Boxplots



ii. P-values of post-hoc Tukey HSD

Scores for language groups were significantly different in both original and updated RMET ($p < .0001$).