Individual Differences in Object Recognition: Supplement

This supplement elaborates on several data-analytic issues and procedures used in our manuscript. The data and code for Study 1 can be found at https://figshare.com/s/24b5c4a510f87c7d6b5a. The data and code for Study 2 can be found at https://figshare.com/s/e962df58149a0a2a3af2.

Intraclass Correlations

To estimate the intraclass correlations (ICCs) and confidence intervals shown in Table 4, we specified a mixed-effects model that predicted scores on each task by fixed (consistency ICC) or random (agreement ICC) effects for category, a random effect for person, and a random residual term (analogous to a person x category interaction in a conventional random effects ANOVA). To generate estimates and confidence intervals, we used a Bayesian procedure instantiated in SAS PROC MCMC. Vague prior distributions were specified for fixed effects (each normal with $\mu = 0$ and $\sigma^2 = 100$), random variance parameters (other than the random effect for persons; log uniform bounded between -10 and 10) and residual variance parameters (log uniform bounded between -10 and 10), and for ICC_1 (uniform between 0 and 1) (Spiegelhalter, 2001). Using Markov Chain Monte Carlo (MCMC) simulation methods with a 10,000 sample burn-in phase, we generated 50,000 samples. We thinned the samples (selecting every 5^{th}) and thus generated 10,000 samples from the posterior distribution of the parameters. For each sample, we algebraically computed estimates of the variance due to persons $\left(\sigma_p^2\right)$ and of ICC_5 from estimates computed in the first stage. We have found that an approach in which ICC_1 is given a prior distribution and the estimates yielded by MCMC algorithm are used to solve algebraically for the random variance parameter for persons yields more stable solutions and

better sampling of the parameter space than an approach specifying priors for all random and residual parameters and solving for ICC_1 after the fact. We computed medians of the posterior distribution as our ICC estimates and formed 95% Bayesian Highest Posterior Density (HPD) intervals that represent the narrowest intervals with 95% probability (e.g., Christensen, Johnson, Brascum, & Hanson, 2011).

Bifactor Models

An alternative hierarchical structure model to the second-order factor model is a bifactor model (e.g., Chen, West, & Sousa, 2006; Reise, 2012). A bifactor alternative to the second-order factor model shown in the bottom panel of Figure 4 would specify direct paths from a general factor (i.e., *o*) to each of the 15 task indicators (3 tasks for each of 5 categories). Five "group" factors would represent the unique variance due to each category on which the 3 task indicators for a given category would load. The intercorrelations among all 6 factors would be constrained to equal 0 to insure an identifiable (i.e., estimable and testable) model. It can be shown that the second-order factor model shown in Figure 4 is actually a restricted version of a bifactor model (e.g., Yung, Thissen, and McLeod, 1999). That is, it imposes the same restrictions on the data as the bifactor model, plus additional ones involving proportionality constraints among factor loadings. A statistical test (i.e., a nested chi-square test, see below) comparing the two models indicated that the additional restrictions imposed by the second-order model could not be rejected ($p > .10$). Other indices of model fit discussed in the Data Analysis section of Study 1 indicated either that the models fit equivalently or that the second-order model fit slightly better, especially when model parsimony (fewer parameters specified) was rewarded. We also specified a second type of bifactor model according to which all tasks load directly on *o* and the group factors are task, rather than category, factors. Although this model fit rather well (e.g., RMSEA =

.044), it fit worse than the second-order factor model when within-task factor loadings were constrained equal as in Model 4 (RMSEA for bifactor model = .067).

Apart from the relative fit of competing models, there are two primary reasons why we focused on the second-order factor model. Simulation studies have raised increasing concern about the fitting propensity of bifactor models. Fit indices can commonly indicate that bifactor models fit better than a correlated factor model or second-order factor model even when one of the latter alternatives represents the correct population structure for the data (e.g., Morgan, Hodge, Wells, & Watkins, 2015; Murray & Johnson, 2013). More generally, there is evidence that bifactor models are over-flexible (i.e., susceptible to overfitting) by, for example, accommodating mis-specifications, random noise, and nonsense response patterns (e.g., Bonifay, Lane, & Reise, 2017; Reise, Kim, Mansolf, & Widaman, 2016).

A second consideration is our belief, shared with other methodologists (e.g., Morgan et al., 2015), that substantive and conceptual considerations should be a major determinant of the decision concerning which model to emphasize. From that perspective, we believe that the second-order factor model that is the focus of our analyses is quite compelling because it posits a causal chain of influence that in the present context makes sense. According to this perspective, a higher-order object recognition ability influences learning and/or recognition of specific categories of objects that in turn influences performance on specific tasks designed to assess ability on each category. Finally, inclusion of additional details might cloud the essential findings and their substantive implications.

While our models showed a strong fit to the observed data, there are issues regarding model fit that any given set of CFA and SEM analyses typically do not and/or cannot address. One of these is the overall fitting propensity of a given model (e.g., Pitt, Myung, & Zhang,

2002), that is, the degree to which it might fit data that arise out of a diverse array of alternative population structures. The greater the fitting propensity, the more flexible a model and the more likely it is to fit even when it does not accurately characterize the true structure generating the data. Assessments of fitting propensity that have been conducted to date do clearly indicate that the higher-order factor model is sensitive to misspecifications and has a significantly more restricted fitting propensity than the bifactor model (e.g., Mansolf & Reise, 2017; Morgan et al., 2015) The assessment of the fitting propensity of SEM models is a complex enterprise (Preacher, 2006) that requires significant methodological development. This will be a target of our future work.

Robust Estimation

As noted in the description of the CFA analyses conducted in Study 1, we used the two-stage Savalei-Bentler two-stage (TS) estimator to deal with the issues of non-normality and the presence of missing data. The TS estimator yields estimates that are consistent (converging in probability to the true parameters with increasing sample size) for non-normally distributed data as long as the mechanism for missing data is either missing completely at random (MCAR) or missing at random (MAR) under the most common conditions for the latter (Yuan & Bentler, 2000; Yuan & Lu, 2008). It should be noted that the Savalei-Bentler TS estimator is not the same as the TS estimator used in earlier simulation studies (e.g., Enders & Peugh, 2004) that fails to correct appropriately standard errors and the model chi-square even under normality (Savalei & Bentler, 2009).

At the present time, robust full-information maximum likelihood (FIML) is more commonly used in SEM studies to analyze data with both non-normality and missing data. We believe that one major reason is the greater availability of this estimator in commonly used

software (e.g., the MLR estimator in MPLUS). We used the Savalei-Bentler approach rather than a robust full-information maximum likelihood (FIML) approach because Savalei and Falk's (2014) results for sample size and missing data conditions that mirrored those of the present study generally indicated that the TS approach performed better. Results were, however, very similar when analyses were conducted with robust FIML using either EQS or MPLUS. We also decided to use the Savalei-Bentler TS estimator rather than the two-stage robust two-state estimator developed by Yuan and colleagues (Tong, Zhang, & Yuan, 2014; Yuan & Zhang, 2012) that down-weights extreme observations because: (1) The Savalei-Bentler TS approach affords more accessible computation of model fit indices beyond the chi-square test of exact fit; (2) We believe that the most appropriate implementation of model comparisons using the Yuan-Zhang estimators is an issue that requires further study;[1] (3) Violations of normality (in particular, kurtosis values) in the present study were relatively moderate and outliers were rare; and, (4) Based on prior simulation results (e.g., Savalei & Falk, 2014; Tong et al., 2014), the percentage of missing data (~15%) was below the fraction (e.g., 30%) that would engender significant concern about the performance of the Savalei-Bentler robust TS estimator. Results were similar when analyses were conducted using the Yuan-Zhang estimators.

Latent Variable Analyses in Study 2.

As noted in the text, due to the sample size (N=54), Study 2 is far from ideal for latent variable modeling using SEM software. For this reason, in the text, we focused on the analyses of observed measures and only discuss SEM analyses for the FIQ latent variable. Because, however, the results of such analyses may be of interest to readers we summarize them in this

[1] For example, we found one case using the Yuan-Bentler residual-based statistic in which the chi-square value for a nested model with more restrictions (greater degrees of freedom) was slightly lower than that for a comparison model that imposed fewer restrictions.

section. For each measure, we specified two models that are similar to the two models for FIQ discussed in the text. The first model was a factor model designed to estimate correlations between the Cat0 and Cat2 latent variables and a latent variable representing a given measure (e.g., visual short-term memory). Instead of using multiple indicators for variables other than FIQ we used latent variables that represented reliability-corrected scores (i.e., models were specified such that the proportion of the total observed variance of a given measure due to the latent factor equaled the estimated reliability of that measure reported in Table 8). The second model was an equivalent model that regressed each of the Category factors on the latent variable denoting a specific individual difference measure and allowed the residuals of Cat0 and Cat2 to be freely correlated.

Given the non-normality of some of the variables, we used robust estimation techniques using MPLUS. When data were missing (1 observation for Stroop Cost and Visual STM) we used the MLR estimator and for the other measures we used the MLM estimator. Both use the Satorra-Bentler correction for standard errors and test statistics. We used MPLUS rather than EQS for these analyses simply because of the greater ease of generating estimates and confidence intervals for the derived measures discussed below and of generating bias-corrected bootstrap confidence intervals. In fact, confidence intervals were almost identical whatever the software platform or resampling procedure and they were very similar across different bootstrap confidence interval approaches (e.g., BCA, bias-corrected, asymptotic adjusted, or percentile). This is due in part to the extremely small percentage of missing data in Study 2. Consistent with the observed variable results presented in Table 9, we computed bias-corrected bootstrap confidence intervals around correlations and partial correlations and percentile confidence intervals around other measures.

Supplemental Table 1 shows the fit of each model and Supplemental Table 2 shows parameter estimates and confidence intervals. Given the sample size, these confidence intervals should be regarded with caution (e.g., Nevitt & Hancock, 2001). All models specified reached convergence and had proper solutions. Table 1 shows only one set of fit indices per measure because the two models specified for each measure are equivalent (i.e., their fit is identical). As shown in Supplemental Table 1, in most cases, the models fit well. Including FIQ, 6 of the 9 models consistently yield values of the RMSEA and other indices that indicate good to excellent fit (RMSEA < .06). The exceptions are Visual Short-term Memory, Shift Cost, and Emotional Stability, although the latter two have RMSEA values that indicate at least reasonable fit. However, readers should note the wide confidence intervals around the RMSEA values. Although confidence intervals were not estimated, the other fit measures undoubtedly have the same feature.

Supplemental Table 2 shows estimates and confidence intervals for the correlational (zero-order and partial) and decomposition measures discussed in the text. Values for Visual STM in particular should be interpreted with caution given the non-optimal fit of its models. Also it should be noted that, while we reliability-corrected all measures, the reliability of Stroop cost was .50 (see Table 8), which suggests that estimates may not be optimally precise. Even give these caveats, it is evident from this table that: (1) Correlations between the individual difference measures and Cat 0 and Cat 2 tend to be modest at best with the possible exception of Shift Cost; (2) Partial correlations between Cat0 and Cat2 tend to be very high and approach the zero-order correlation between Cat0 and Cat2 (r=.89); and, (3) Negligible to small components of the overall correlation between Cat0 and Cat2 go through the individual difference variables. Conversely, as indicated by the last column, the lowest value of the proportion due to the

residual path is 83% (Shift Cost) and most values of this index are in the 90-100% range. An additional feature of Supplemental Table 2 is that, despite the sample size, many of the confidence intervals are relatively narrow. For example, among the individual measures, the lowest lower bound for a 95% confidence interval around the proportion of residual component measure is 56% and the lowest lower bound for the correlation among the residuals is .61 Considered as a whole, these results are consistent with the results summarized in the text that focus on observed correlations and the latent variable model for FIQ.

References

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, 1-20.

Bonifay, W., Lane, S.P., & Reise, S.P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science, 5*, 184-186.

Chen, F.F., West, S.G., & Sousa, K.H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*, 189-225.

Christensen, R., Johnson, W., Branscum, A., & Hanson, T. E. (2011). *Bayesian ideas and data analysis: an introduction for scientists and statisticians*. Boca Raton, FL: CRC Press.

Enders, C.K., & Peugh, J.L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. *Structural Equation Modeling, 11,* 1-19.

Mansolf, M., & Reise, S.P. (2017). When and why second-order and bifactor models are distinguishable. *Intelligence, 61*, 120-129.

Morgan, G. B., Hodge, K. J., Wells, K. E., & Watkins, M. W. (2015). Are fit indices biased in favor of bi-factor models in cognitive ability research?: A comparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. *Journal of Intelligence, 3*(1), 2-20.

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, *41*(5), 407-422.

Nevitt, J., & Hancock, G.R. (2001). Performance of bootstrapping approaches to model tests statistics and parameter standard error estimation in structural equation modeling.

Structural Equation Modeling, 8, 353-377.

Pitt, M.A., Myung, I.J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472-491.

Preacher, K.J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research, 41,* 227-259.

Reise, S.P. (2012). Invited paper: The rediscovery of bifactor measurement models. *Multivariate Behavioral Resarch, 47,* 667-696.

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research, 51*(6), 818-838.

Savalei, V., & Bentler, P.M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling, 16,* 477-497.

Savalei, V., & Falk, C.F. (2014). Robust two-stage approach outperforms robust full information maximum likelihood with incomplete nonnormal data. *Structural Equation Modeling, 21,* 280-302.

Spiegelhalter, D. J. (2001). Bayesian methods for cluster randomized trials with continuous responses. *Statistics in Medicine, 20*(3), 435-452.

Tong, X., Zhang, Z., & Yuan, K-H. (2014). Evaluation of test statistics for robust structural equation modeling with nonnormal missing data. *Structural Equation Modeling, 21,* 553-565.

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30,* 165–200.

Yuan, K-H., & Lu, L. (2008). SEM with missing data and unknown population using two-stage ML: Theory and its application. *Multivariate Behavioral Research, 62,* 621-652.

Yuan, K-H., & Zhang, Z. (2012). Robust structural equation modeling with missing data and auxiliary variables. *Psychometrika, 77(4),* 803-826.

Yung, Y-F., Thissen, D., & McLeod, L.D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64,* 123-128

Supplemental Table 1: Measures of Fit for Study 2 Models

Measure	RMSEA	CFI	SRMR
Stroop Cost	.019 (.000,.139)	.998	.042
Shift Cost	.068 (.000,.163)	.982	.052
L-EFT	.058 (.000,.147)	.984	.044
Visual STM	.122 (.036,.199)	.927	.054
Conscientiousness	.040 (.000,.143)	.984	.046
Extraversion	.059 (.000,.150)	.984	.044
Emotional Stability	.086 (.000,168)	.967	.050
Agreeableness	.000 (.000,.137)	1.00	.041
Intellect	.046 (.000,.148)	.991	.051

Note. N=54. 90% confidence intervals are shown for the RMSEA using the Li and Bentler (2006) formula. Hu and Bentler (1998, 1999) recommended the following criteria for adequate fit on the first three measures: CFI ≥ .95, RMSEA ≤ .06, and SRMR ≤ .08.

Supplemental Table 2: Zero-order, Partial, and Decomposed Correlations for Latent Variables in Study 2

Individual Difference Measure	Cat 0 r	Cat 2 r	Cat0/Cat2 Partial r	Component of Cat0/Cat2 r Through Ind. Dif. Path	Component of Cat0/Cat2 r Through Residual Path	Proportion: $\left(\frac{\text{Residual Component}}{\text{Cat0/Cat2 r (= .88)}}\right)$
Stroop Cost	-.22 (-.61.10)	-.30 (-.63,.05)	**.87** (.61,.1.00)	.07 (-.01,.39)	**.81** (.45,.95)	**92%** (56%,101%)
Shift Cost	**-.45** (-.68,-.15)	**-.33** (-.56,-.09)	**.86** (.63,1.00)	**.15** (.02,.37)	**.73** (.46,.92)	**83%** (57%,98%)
L-EFT	.11 (-.11,32)	-.02 (-.30,.24)	**.88** (.67,.1.00)	-.00 (-.02, .06)	**.87** (.66,.1.01)	**100%** (93%,102%)
Visual STM	.32 (-.00,.59)	.35 (-.01,.58)	**.87** (.62,1.00)	**.11** (.00,.32)	**.77** (.51,.94)	**87%** (63%,100%)
Conscientiousness	-.24 (-.53,.10)	-.19 (-.51,.15)	**.88** (.67,1.00)	. 05 (-.01,.26)	**.84** (.58.99)	**95%** (71%,100%)
Extraversion	-.02 (-.26,.23)	.00 (-.31,.26)	**.88** (.68,1.00)	.00 (-.01,.07)	**.88** (.68,1.00)	**100%** (92%,102%)
Emotional Stability	-.11 (-.35.18)	-.00 (-.30,.28)	**.88** (.67,.1.00)	.00 (-.02,.08)	**.87** (.64,.1.00)	**100%** (88%,102%)
Agreeableness	-.16 (-.48,.16)	-.08 (-.38,.20)	**.88** (.68,1.00)	.01 (-.01,.17)	**.86** (.64,1.00)	**99%** (81%,101%)
Intellect	-.03 (-.35,.27)	.20 (-.15,.44)	**.89** (.72,.1.00)	-.01 (-.05,.11)	**.87** (.68,.1.00)	**101%** (88%,107%)

Note. Zero-order correlation between Cat0 and Cat2 latent variables = .89. Partial correlations between Cat0 and Cat2 adjust for the given individual difference measure. Component indices decompose the zero-order correlation between Cat0 and Cat2 into components through the individual difference measure and the residual paths. Bias-corrected bootstrap confidence intervals are shown in parentheses for zero-order and residual correlations. Percentile bootstrap confidence intervals are shown for the final three indices. Estimates are bolded when confidence intervals do not include 0. Upper bounds for proportions in the last column can exceed 1 due to negative values for the component of the correlation through the individual difference path.