

Intelligence Test Items varying in Difficulty cannot be used to test the Causality of Working  
Memory Capacity for Intelligence

Gidon T. Frischkorn & Klaus Oberauer

University of Zurich, Switzerland

**Author Note**

Add author note, if applicable.

### Abstract

It is well-established that intelligence and working memory capacity are closely related. The cognitive mechanisms underlying this relationship are, however, still under debate. One popular hypothesis, the *capacity hypothesis*, states that this relationship is caused by limitations in the amount of information that can be stored and held active in working memory. Previous research testing this hypothesis assumed that the capacity hypothesis implies stronger relationships of more difficult intelligence test items, or items requiring to maintain more information (e.g. sub-goals), with measures of working memory capacity. The present article addresses this assumption in a simulation systematically varying different psychometric variables: the mean sample ability, the variability of ability in the sample, item difficulty, and item discrimination. All simulations assumed a single latent ability – which could be working-memory capacity – underlying performance in the test items. The results of these simulations show that almost any relation between item difficulty and their correlation with the latent ability can be obtained. Therefore, the assumption made by previous studies does not hold, and items varying in difficulty cannot be used to test the causality of working memory capacity (or any other latent variable) for intelligence.

*Keywords:* Intelligence, Working Memory Capacity, Item-response theory

Item Difficulty of Intelligence Test Items cannot be used to test the Causality of Working  
Memory Capacity for Intelligence

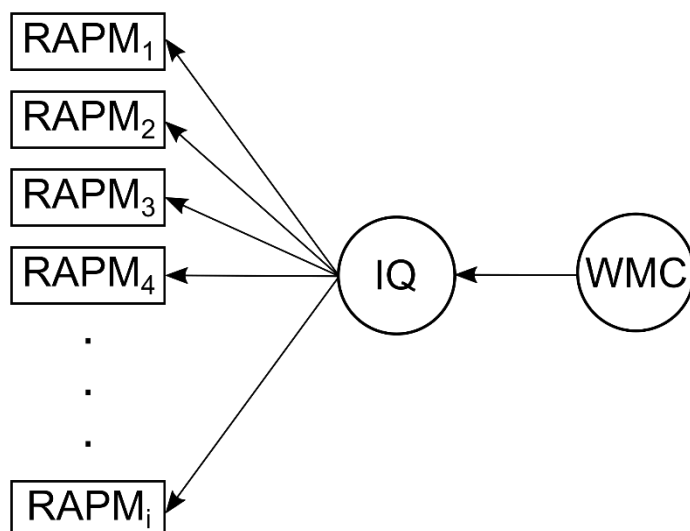
A plethora of research has established that individual differences in intelligence are closely related to measures of working memory capacity (Conway & Kovacs, 2013; Kyllonen & Christal, 1990; Oberauer, Schulze, Wilhelm, & Süß, 2005). What cognitive mechanisms underlie this relationship? A prominent account is that individual differences in working memory capacity (WMC) are causal for intelligence differences. This so-called *capacity hypothesis* assumes that an individual's ability to maintain information in working memory is determining the performance in measures of intelligence (Unsworth, Fukuda, Awh, & Vogel, 2014).

Previous studies claimed that this hypothesis can be tested by comparing the correlation of WMC measures with intelligence test items of varying difficulty (Little, Lewandowsky, & Craig, 2014; Salthouse, 1993; Unsworth & Engle, 2005; Wiley, Jarosz, Cushen, & Colflesh, 2011) or requiring to store varying amounts of information (Burgoyne, Hambrick, & Altmann, 2019). The underlying assumption of these studies was that more difficult items would pose higher demands on WMC and therefore should show higher correlations with measures of WMC. In detail, these studies quantified the correlation of different items from the popular Raven matrices (Raven & Raven, 2003) with measures of WMC, such as span tasks (Daneman & Carpenter, 1980) or visual arrays tasks (Luck & Vogel, 1997). In these analyses, all but one study (Little et al., 2014) found no increase in correlations with WMC for more difficult Raven items. These results were interpreted as evidence against a causal role of WMC for intelligence,

indicating that other processes such as attention control (Burgoyne et al., 2019; Wiley et al., 2011) may be more important for intelligence differences.

### Psychometric problems of the derived hypothesis

From a psychometric perspective the critical question is: what does the *capacity hypothesis* – stating that individual differences in WMC are causally related to variation in intelligence test performance – imply for the relationship of WMC with intelligence test items of varying difficulty? The theoretical model describing the capacity hypothesis states that a single ability (i.e. WMC) is causally responsible for individual differences in the performance of intelligence test items. This model is illustrated in Figure 1 (see below) as a path diagram. To test what this model implies for the relationship between WMC and performance on the intelligence test items (RAPM<sub>1</sub> to RAPM<sub>i</sub>), we can systematically vary the item characteristics of the intelligence test items and investigate how the relationship of performance on the test items with WMC looks like for different item characteristics.



**Figure 1.** Path diagram of a theoretical model illustrating the capacity hypothesis. Working memory capacity (WMC) causally determines intelligence (IQ) that is in turn responsible for the performance in different intelligence test items (RAPM<sub>1</sub>, RAPM<sub>2</sub>, RAPM<sub>3</sub>, RAPM<sub>4</sub>, ..., RAPM<sub>i</sub>). Circles are used to illustrate latent, not directly observable constructs. Rectangles are used to illustrate manifest performance observed as the accuracy of a response in an intelligence test item.

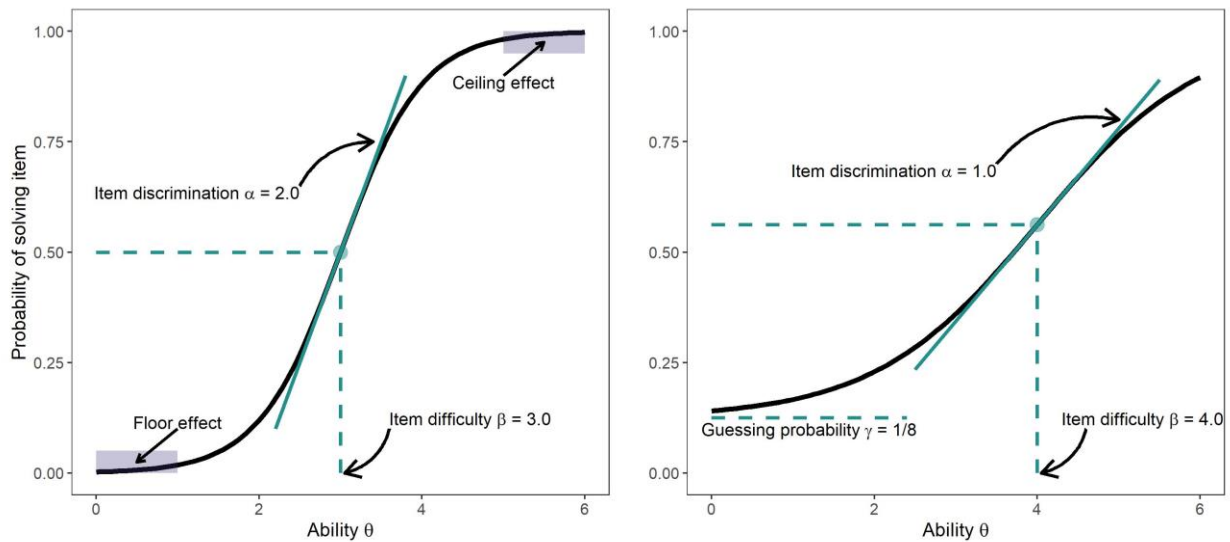
In psychometrics two item characteristics are commonly distinguished: a) *item difficulty*, which captures how many people are able to solve an item, and b) *item discriminability* that represents how well an item can separate people varying in the to be measured ability. Specifically, these two characteristics are related to two different measurement properties of an item: *item difficulty* determines at which ability level an item will differentiate, and *item discrimination* determines how well it will do so. Originally (Lord & Novick, 1968), these item characteristics were defined as statistical variables calculated within a sample: *item difficulty* as the proportion of correct responses to an item, and *item discrimination* as the correlation of a single item with the total score.

With this definition of *item difficulty* and *discrimination*, the estimates for these item characteristics depend on the average level and distribution of the ability in the sample used for calculation. For instance, if you consider a high ability sample, more people will be able to solve test items in this sample and thus *item difficulty* (i.e. the proportion of correct responses) will be estimated lower than in a sample with lower overall ability. Similarly, the correlation of item performance in easy items with the total test score (i.e. *item discrimination*) will be lower in a high ability sample due to ceiling effects in item performance, whereas item discrimination estimates for the same item will be higher in a low ability sample. In summary, with this definition of *item difficulty* and *discrimination*, estimates for these item characteristics are not only properties of an item but also depend on sample characteristics.

This sample dependency has actually been discussed with respect to diverging results regarding the capacity hypothesis: Little et al. (2014) found almost ceiling performance for early (i.e., easy) Raven items whereas other studies (Burgoyne et al., 2019; Unsworth & Engle, 2005;

Wiley et al., 2011) did not. In consequence, Little et al. (2014) found stronger correlations for more difficult Raven items with WMC whereas the other studies found equal correlation for items of varying difficulty with WMC. Despite the known sample dependency of item characteristics, this was rather seen as a problem regarding the generalizability and validity of Little et al.'s (2014) results (Burgoyne et al., 2019) than for what it would be if the *capacity hypothesis* were true: A psychometric necessity.

An alternative to the initial definition of item difficulty and discrimination that overcomes the problem of sample dependency of these item characteristics is offered by item-response theory (Birnbaum, 1968; Rasch, 1993). Item-response theory (IRT) models provide a separation of different parameters that may affect the performance of a person on a specific test item. In detail, the probability that a person can solve a test item primarily depends on their ability  $\theta_p$  and the difficulty  $\beta_i$  of the item. This relationship is best described by item-response functions (IRF) that are illustrated in Figure 2. At the core, the probability to solve an item is determined by the difference between ability  $\theta_p$  and item difficulty  $\beta_i$ . If the ability is larger than the item difficulty then a person is likely to solve an item, whereas a person is unlikely to solve an item when their ability is lower than the item difficulty. The strength of this effect is additionally determined by the item discrimination  $\alpha_i$  (illustrated by the slope in Figure 2). If item discrimination is high (see left side of Figure 2) changes in ability (e.g. from  $\theta_p = 2.5$  to  $\theta_p = 3.5$ ) have large effects on the probability to solve an item. In contrast, if item discrimination is low (see right side of Figure 2) the same change in ability will have a smaller effect on the probability to solve an item. Finally, the lowest possible probability to solve an item is captured in the guessing parameter  $\gamma_i$ , illustrated by the asymptote on the right side of Figure 2. Formally,



**Figure 2.** Illustration of two item-response functions (IRF) for two items with varying difficulty  $\beta$ , discrimination and guessing probability. The IRF on the left corresponds to an easy item with high discrimination. The IRF on the right corresponds to a difficult item with low discrimination and a guessing probability of  $1/8$ . Although high item discrimination is better able to differentiate between people high and low in ability around its location (i.e. item difficulty), both floor and ceiling effects – as illustrated on the left side – can occur for items with high discrimination.

the probability for solving an item in such a three-parameter logistic item response model is

defined as:

$$P_i(\theta) = \gamma_i + (1 - \gamma_i) \frac{e^{\alpha_i(\theta_p - \beta_i)}}{1 + e^{\alpha_i(\theta_p - \beta_i)}}$$

In empirical applications the estimation of the different parameters can still depend on the ability of the sample (mean and variability) as well as the selection of items. In particular, item discrimination can be estimated best around the region where the ability matches an item's difficulty (i.e.  $\theta_p - \beta_i = 0$ ) whereas item discrimination estimates will be unreliable the farther ability is above or below an item's difficulty (see the left side of Figure 2 for these floor and ceiling effects that bear little information regarding the slope of the item-response function). In a simulation however, the ability of the sample can be varied independently from item difficulties and discriminations. Thus, using an IRT simulation can illustrate the implications of the *capacity*

*hypothesis* for the correlation of items with varying difficulty or discrimination and the underlying ability (e.g. WMC) independent of sample characteristics that may affect the results in an empirical study.

### **The present IRT simulation**

Previous studies assumed that the *capacity hypothesis* stating that WMC causally underlies individual differences in intelligence test performance implies increasing correlations for more difficult items, or items that require to store more information, with measures of WMC. The present simulation investigated in how far this prediction can be derived from the *capacity hypothesis*. This is essential to evaluate the results of previous studies and their interpretation regarding the relationship between WMC and intelligence. If the hypothesis motivating previous studies is true, a model assuming WMC to be the common cause of intelligence test performance should produce systematically higher correlations with WMC for more difficult items, or items requiring more information to be stored in WM. In contrast, if this hypothesis was false, then a multitude of different correlation patterns across items with varying difficulty could arise that are all in line with the capacity hypothesis.

## **Methods**

### **IRT simulation**

To align the present simulation (R scripts available at: [osf.io/rt2j8](https://osf.io/rt2j8)) with the most recent publication that aimed at testing the capacity hypothesis with Raven items (Burgoyne et al., 2019), we simulated data for the 17 intelligence test items used in their study. Following the

rationale of Burgoyne et al. (2019) the difficulty of each item was specified depending on the number of rule tokens the respective item required ( $\beta_i = 1, 2, 3, \text{ or } 4$ ) according to analysis of Raven items by Carpenter et al. (1990). Then, the latent ability for 250 subjects was drawn from a truncated Gaussian distributed (Min = 0, Max =  $\infty$ ) for which mean ability level ( $\mu_\theta$ ) as well as standard deviation of the ability ( $\sigma_\theta$ ) varied randomly ( $\mu_\theta = [0.2; 5]$ ;  $\sigma_\theta = [0.25; 2]$ ) for each simulation run.

In addition, to investigate the role of item discrimination, different conditions were run assuming item discrimination to be equal, random, increasing, or decreasing across item difficulties. In each of these conditions, item discrimination ( $\alpha_i$ ) was randomly drawn from a uniform distribution ( $\alpha_i = [0.5, 4]$ ). For equal item discrimination across item difficulties one value was drawn and used for all items. For random, increasing, and decreasing item discriminations across item difficulties, four values were drawn – one for each difficulty level – and sorted accordingly for the respective conditions. Each of these item discrimination conditions was repeated 1000 times resulting in a total of 4000 simulations. The guessing probability  $\gamma_i$  was held constant at 1/8 for all simulation runs as Raven items force participants to choose the correct solution from 8 options.

Finally, to address the question whether the pattern of correlations of single items with WMC depends on the overall correlation between intelligence and WMC, the simulation was run assuming that the ability underlying the test item performance (IQ) was not identical to WMC but correlated with it at different levels ( $\rho = .90, .60 \text{ or } .30$ ). In total the 4000 simulation runs were thus repeated for each of the three levels of correlation between the latent ability and WMC.

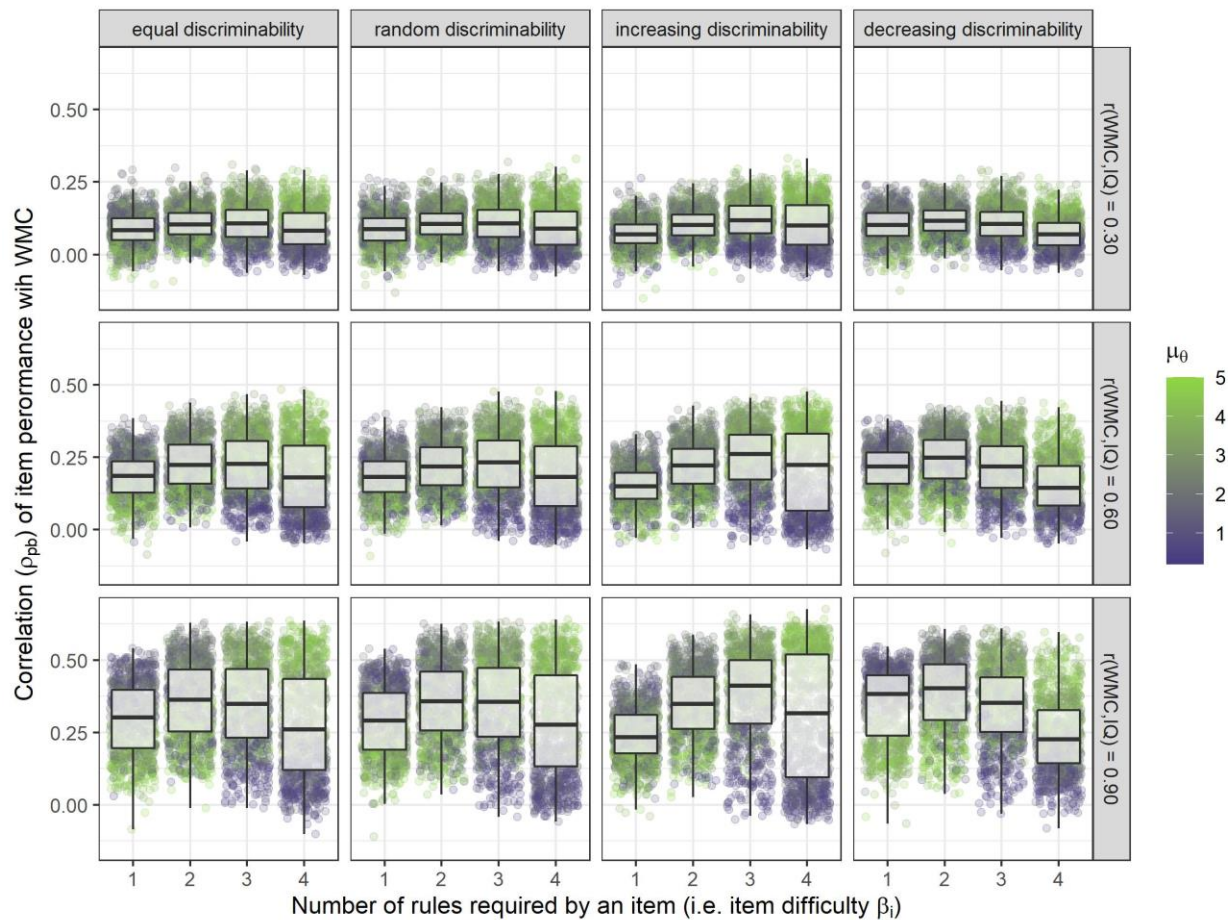
Within each simulation the probability of solving an item correctly was determined according to the randomly drawn person ability and the item difficulties of the 17 Raven items as specified by the number of rule tokens (see Table 3 in Burgoyne et al., 2019). This probability was then used to randomly draw from a binomial distribution whether a person solved an item correctly or not.

### **Evaluation of simulation results**

To evaluate the effects of different simulation conditions on the pattern of correlations across different item conditions, we plotted the resulting point-biserial correlations of item performance with WMC across the number of rules required by an item, and the other simulation conditions. As the strength of evidence for or against the effect of specific conditions is merely a function of the number of simulations, we refrained from computing any statistical tests. Instead, we descriptively evaluated these plots and verbally summarized the results.

### **Results**

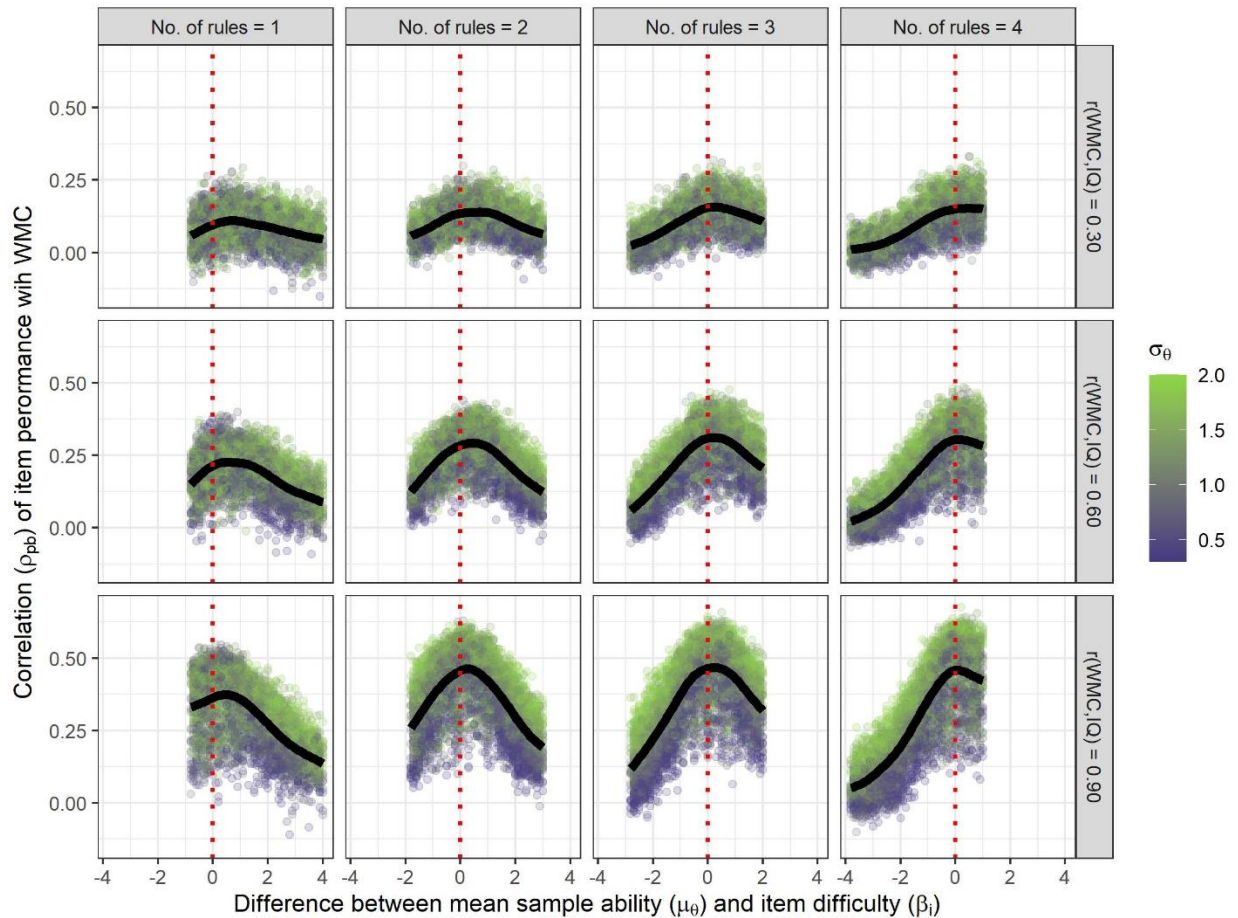
Analogous to results from Burgoyne et al. (2019), Figure 3 illustrates the pattern of correlations across items requiring different number of rules with WMC. The main takeaway from this figure is that although we assumed a single latent ability (IQ) – solely predicted by WMC – underlying performance in all items, the pattern of correlations of item performance with WMC was not consistently increasing from easy to difficult items. On the contrary, the results indicate that the correlations can either decrease, increase or remain the same across different item difficulties. The pattern mainly depended on the way item discrimination is changing with item difficulty. If discrimination increases with difficulty (third column in Figure 3) then



**Figure 3.** Pattern of point-biserial correlations (colored dots) averaged across the number of rules required by an item (i.e. item difficulty). The boxplots summarize the median, inter-quartile range and total range across all simulations. The rows separate results assuming different strength of correlation between WMC and the intelligence ability (IQ) underlying item performance. The columns separate results for different patterns of item discrimination across item difficulty. Finally, the color coding indicates the mean of the ability underlying item performance.

correlations increase from easy to difficult items. If discrimination decreases with item difficulty (fourth column in Figure 3) then correlations decrease from easy to difficulty items. Thus, rather than item difficulty, item discrimination determines the correlation of item performance with WMC.

In addition, the correlation of performance across items with varying difficulty strongly depends on the mean level of ability ( $\mu_\theta$ ). As shown by the color coding in Figure 3, easy items show strong correlations when mean ability is low (dark violet dots), whereas difficult items



**Figure 4.** Correlation of item performance with WMC plotted depending on the difference of mean sample ability and item difficulty. The vertical dotted red line indicates a perfect match between sample ability and item difficulty (i.e.  $\mu_{\theta} - \beta_i = 0$ ), and the black line represents the average pattern of correlations fitted with a generalized additive model. The color coding of the dots indicates the variability of the ability in the respective simulation run.

show strong correlations when ability is high in the sample (light green dots). This becomes even more obvious when plotting the correlation of item performance with WMC depending on the difference between average sample ability and item difficulty (see Figure 4). This shows that the correlation of performance with WMC is always maximal when the average sample ability closely matches the difficulty of an item (i.e.  $\mu - \beta = 0$ ) independent of its overall difficulty

(illustrated by the different columns). In addition, the color coding in this plot shows that the correlation of item performance with WMC increases the larger variability of the ability is.

In sum, the results from this simulation show that assuming a theoretical model in line with the *capacity hypothesis* does not imply that correlations with WMC increase with item difficulty. Instead, the degree to which the ability of the sample matches the difficulty of an item seems to determine the correlation. Beyond that, the correlation of item performance with the underlying ability increases as item discriminations, and the variability of the ability in the sample, increase. Taken together, these effects result in increasing, decreasing, or constant patterns of correlations with WMC across items with varying difficulty. Therefore, contrary to the assumptions underlying previous studies, the capacity hypothesis does not imply that correlations between difficult items and WMC should be larger than between easy items and WMC.

### **Discussion**

The aim of the presented simulation study was to assess what the *capacity hypothesis* implies for the pattern of correlation of items varying in difficulty, or requiring different numbers of rules to be maintained in memory with WMC. Unlike what previous studies assumed (Burgoyne et al., 2019; Little et al., 2014; Salthouse, 1993; Unsworth & Engle, 2005; Wiley et al., 2011), increases in item difficulty do not imply an increase in correlation between items with increasing difficulty and WMC. Although this hypothesis appears plausible, it neglects the difference between *item difficulty*, representing at which level of an ability an item is a useful

measure, and *item discrimination*, representing how well an item separates people high and low in ability at the level determined by item difficulty (Birnbaum, 1968; Rasch, 1993).

The results of the simulation study show that increases in item discrimination, and the variability of the ability, lead to stronger correlations between item performance and WMC independent of item difficulty and the average level of ability in the sample. In contrast, the latter two variables influence the correlation only depending on one another. Specifically, the correlation between item performance and WMC was largest in the simulation when the average ability of the sample matched the difficulty of an item.

**Do previous results and the presented simulation confirm the *capacity hypothesis*?**

It would be premature to conclude from the presented simulation that results of previous studies confirmed the capacity hypothesis. Instead, the results of previous studies are consistent with the capacity hypothesis, and hence do not falsify it. The capacity hypothesis does not appear to make any specific prediction regarding the pattern of correlations across items varying in difficulty and WMC. Therefore, the conclusions drawn from previous studies (Burgoyne et al., 2019; Unsworth & Engle, 2005; Wiley et al., 2011) that not working memory capacity but other cognitive processes such as attention control are more relevant for intelligence differences is not warranted.

The present simulation results are not limited to WMC as the common cause of intelligence differences. The results hold for any single latent variable, such as attention control, assumed to determine performance differences on intelligence test items. Therefore, the analysis of correlation patterns across intelligence test item of varying difficulty with any indicator of a

hypothetical cause of intelligence will not be informative regarding its causality. Instead more elaborate and theoretically-grounded measures for specific cognitive processes (Frischkorn & Schubert, 2018), or experimental studies that use manipulations that ideally target a single cognitive process (Rao & Baddeley, 2013; Schubert, Hagemann, Frischkorn, & Herpertz, 2018) are needed to investigate the cognitive processes underlying intelligence differences.

### **Open Practices Statement**

To ease the reproducibility of both the simulation and all results, the R script running the simulation as well as generating all figures included in this manuscript can be accessed via [osf.io/rt2j8](https://osf.io/rt2j8).

### **References**

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examiner's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading: Addison-Wesley.
- Burgoyne, A. P., Hambrick, D. Z., & Altmann, E. M. (2019). Is working memory capacity a causal factor in fluid intelligence? *Psychonomic Bulletin & Review*, *26*, 1333–1339.
- Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *37*, 404–431.

Conway, A. R. A., & Kovacs, K. (2013). *Chapter Seven—Individual Differences in Intelligence and Working Memory: A Review of Latent Variable Models* (B. H. Ross, Ed.). In (pp. 233–270). Academic Press.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

Frischkorn, G. T., & Schubert, A.-L. (2018). Cognitive Models in Intelligence Research: Advantages and Recommendations for Their Application. *Journal of Intelligence*, *6*, 34.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*, 389–433.

Little, D. R., Lewandowsky, S., & Craig, S. (2014). Working memory capacity and fluid abilities: The more difficult the item, the more more is better. *Frontiers in Psychology*, *5*.  
<https://doi.org/10/gfpgnc>

Lord, F. M., & Novick, M. R. (1968). *Statistical theoreis of mental test scores*. Reading: Addison-Weseley.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.-M. (2005). Working Memory and Intelligence--Their Correlation and Their Relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*, 61–65.

Rao, K. V., & Baddeley, A. (2013). Raven's matrices and working memory: A dual-task approach. *The Quarterly Journal of Experimental Psychology*, *66*, 1881–1887.

- Rasch, G. (1993). *Probabilistic Models for Some Intelligence and Attainment Tests*. MESA Press, 5835 S.
- Raven, J., & Raven, J. (2003). Raven Progressive Matrices. In R. S. McCallum (Ed.), *Handbook of Nonverbal Assessment* (pp. 223–237). Boston, MA: Springer US.
- Salthouse, T. A. (1993). Influence of working memory on adult age differences in matrix reasoning. *British Journal of Psychology*, *84*, 171–199.
- Schubert, A.-L., Hagemann, D., Frischkorn, G. T., & Herpertz, S. C. (2018). Faster, but not smarter: An experimental analysis of the relationship between mental speed and mental abilities. *Intelligence*, *71*, 66–75.
- Unsworth, N., & Engle, R. W. (2005). Working memory capacity and fluid abilities: Examining the correlation between Operation Span and Raven. *Intelligence*, *33*, 67–81.
- Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology*, *71*, 1–26.
- Wiley, J., Jarosz, A. F., Cushen, P. J., & Colflesh, G. J. H. (2011). New rule use drives the relation between working memory capacity and Raven’s Advanced Progressive Matrices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 256–263.