Working memory capacity and fluid abilities: The more difficult the item, the more more is better

Daniel R. Little (daniel.little@unimelb.edu.au)
School of Psychological Sciences, The University of Melbourne
Parkville, VIC 3010 Australia

Stephan Lewandowsky (stephan.lewandowsky@uwa.edu.au)
School of Experimental Psychology, University of Bristol
Bristol, UK BS81TU
School of Psychology, The University of Western Australia
Crawley, WA 6009 Australia

Stewart Craig (craig03@student.uwa.edu.au)
School of Psychology, The University of Western Australia
Crawley, WA 6009 Australia

Abstract

Recent evidence has suggested that the relationship between a test of fluid intelligence, Raven's Progressive Matrices, and working memory capacity (WMC) may be invariant across difficulty levels of the Raven's items. We show that this invariance can only be observed if the overall correlation between Raven's and WMC is low. We demonstrate that by using a composite measure of WMC, which yields a higher correlation between WMC and Raven's than reported in previous studies, that there was a significant positive relationship between Raven's item difficulty and the extent of the itemwise correlation with WMC. This result puts strong constraints on theories of reasoning and challenges some existing views. **Keywords:** Raven's Progressive Matrices; Working Memory Capacity.

Introduction

Working memory (WM), the architecture responsible for the retention and manipulation of information over short periods of time, is a core component of human cognition. People's working-memory capacity (WMC) shares around 50% of the variance with general fluid intelligence (Kane, Hambrick, & Conway, 2005) and is predictive of performance in a number of reasoning tasks and other measures of higher cognitive ability. However, there is some dispute about the exact nature of the relationship between WMC and one important assay of fluid intelligence, Raven's Progressive Matrices (e.g., Raven, Raven, & Court, 1998).

Raven's test is designed such that items differ considerably in difficulty, with easy items—presented early in the test—solvable by more than 90% of participants and the hardest items—presented last—being solvable by fewer than 10% of participants. Carpenter, Just, and Shell (1990) presented a taxonomy of rule types that were used to create each of the Raven's items. To illustrate, Figure 1 presents two sample Raven's-like problems created using different rules. The matrix in panel A contains an incremental rule (i.e., the dots increase across items) and a distribution of 3, permutation rule (i.e,. objects with 1, 2 and 3 triangles are permuted across rows and columns). The matrix in panel B contains a constant rule (i.e., the center dot appears in all items) and a distribution of 2 (or logical XOR) rule (i.e., features which appear in the first two objects do not appear in the third object and features which appear only in one of the first two objects also appear in the third object). Carpenter et al.'s rule taxonomy also included feature decrements between objects, logical disjunction rules (OR) and logical conjunction rules (AND). Participants must infer these rules from the objects in the matrix and then predict and select the missing lower right object in the matrix from the set of possible response options.

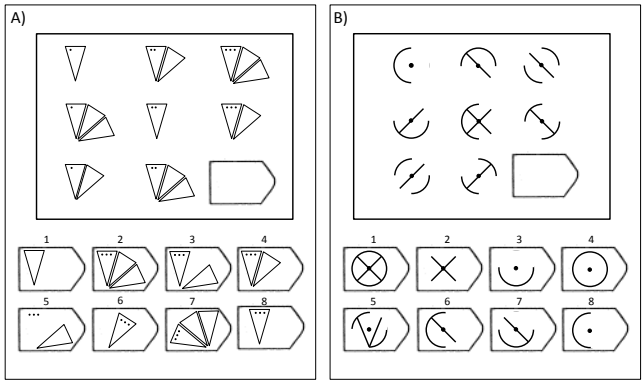

Figure 1: Two examples of matrices like those in the Raven's test. A: Example of an item containing a pairwise incremental rule and a distribution of 3 permutation rule. B: Example of an item containing a constant rule and a distribution of 2 (XOR) rule.

Carpenter et al. (1990) compared two production system models that demonstrated the importance of the number and type of rules and WMC. Both of the models operated by finding correspondences between the

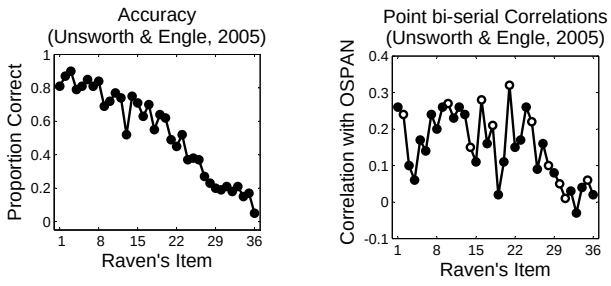

Figure 2: Left Panel: Observed accuracy from Unsworth & Engle (2005). Right Panel: Observed itemwise point bi-serial correlations in Unsworth & Engle (2005).

symbolically-coded features of the items, transferring these correspondences to a working memory buffer where any rule satisfied by the extracted correspondences was invoked, using the instantiated rules to generate the missing item, and finally, searching through the response options to find the best match. One model (called FAIRAVEN) had no strategic memory organization and did not have access to distribution of 2 (XOR) rules; the other model (called BETTERAVEN) was endowed with better control processes and contained access to all of the rules types. The assumptions of the models were consonant with observed accuracy, response time, and eye fixation data and the models were able to explain the performance of median Raven's performers and the very best Raven's performers, respectively.

If we assume that increased WMC allows for an improved ability to maintain goals and retain intermediate results and rules necessary to successfully solve the most difficult Raven's items, the implication of the modeling is that performance on more difficult items should be more highly correlated with WMC. In subsequent tests of that idea, Unsworth and Engle (2005) and Wiley, Jarosz, Cushen, and Colflesh (2011) examined the correlation between WMC and Raven's performance across ordinal item position, which is a proxy for item difficulty. Contrary to expectation, those studies found that the role of WMC remained invariant across item position. The left and right panels of Figure 2 show the accuracy and itemwise correlations observed by Unsworth and Engle (2005). Although the itemwise pattern is quite noisy, there appears to be no systematic relationship between ordinal item difficulty (on the abscissa) and the correlation between performance on those items and WMC (as measured by OSPAN). This impression of an invariant relationship was statistically supported by the failure to find an increasing correlation between OSPAN and the proportion correct within each quartile of the Raven's test.

Those reports of invariant itemwise correlations have been used to reject the model of Carpenter et al. (1990), or indeed any other proposal that cites the ability to hold rules and goals in working memory as underlying Raven's performance. The failure to find a selective involvement of WMC has motivated alternative theorizing about the relationship between the Raven's test and WMC. For example, Unsworth and Engle (2005) concluded that the variance shared by WMC and Raven's reflected attentional control mechanisms, presumed to be implicated in both tasks, which were thought to be uniformly important across all of the Raven's items. Thus, irrespective of item difficulty, a person with larger WMC benefits from an enhanced ability to selectively focus on those features of an item that are relevant to the item-appropriate rule and to filter out distracting non-relevant goals and features. Although this account has not been quantitatively formalized, there is empirical support from other domains that working memory underwrites an ability to filter out distracting information (Conway, Cowan, & Bunting, 2001).

The current state of affairs thus presents a conceptual puzzle: On the one hand, intuition and at least one theory (Carpenter et al., 1990) suggest that the importance of WMC should be accentuated for the more difficult Raven's items, for the simple reason that the easiest items are—by design—solvable by most participants and hence ought not to correlate much with WMC. On the other hand, there are now several reports that the role of WMC is invariant across item difficulty (Unsworth & Engle, 2005; Wiley et al., 2011). Those results appear consonant with an attentional view of working memory but run counter to the model of Carpenter et al. (1990). However, there are several reasons to examine those reports further: First, the counter-intuitive nature of those results deserves to be underscored—after all, how can the correlation between WMC and performance be identical for items that are solved by 90% and 10%, respectively, of participants?

There are other reasons to expect that the acceptance of an invariant relationship between Raven's performance and WMC may have been premature. By definition, those results rely on a failure to reject the null hypothesis, and the "noisiness" of the data is considerable (see Figure 2, right panel). Moreover, studies showing an invariant itemwise correlation were marred by the fact that only a single task (OSPAN) was used to measure WMC—consequently, measurement error or "method variance" from that single task might have masked a relationship between WMC and the more difficult Raven's items in the studies of Unsworth and Engle (2005) and of Wiley et al. (2011). In support of this claim, the correlations reported in those papers ($r = .335$ and $r = .33$, respectively) fall on the lower end of the range of correlations between WMC and Raven's identified in a recent meta-analysis (i.e., .312 to .641; Ackerman, Beier, & Boyle, 2005). Further, Unsworth and Engle (2005), participants were allocated 30 minutes to complete the

Raven's test rather than the standard 40 minutes, which likely resulted in decreased overall accuracy, that may have further obscured an increasing effect of WMC.

We suggest that there are strong and well-supported reasons to expect the involvement of WMC in performance to increase across item difficulty in the Raven's test. Reports to the contrary have relied on acceptance of the null hypothesis and have involved limited measures of WMC. The issue of how working memory relates to Raven's performance may therefore be worthy of further exploration. We revisit this issue and resolve it by presenting a behavioral study using a composite measure of WMC that correlates more strongly with Raven's and results in an increasing itemwise correlation—as predicted by Carpenter et al. (1990) and contrary to the null results reported to date.

Behavioral Study

In this study, we sought to maximize the likelihood of finding an increasing itemwise corrrelation function by using multiple tasks and deriving a composite latent measure of WMC, thus reducing the task-specific variance and measurement error that beset a single-task measure such as OSPAN. We therefore expected the correlation between WMC and RAPM performance to be greater than in relevant previous studies. Why should we expect the overall correlation between WMC and Raven's to affect the itemwise correlation? The answer lies in the constraints imposed by the decreasing accuracy function across Raven's items: Because nearly everyone gets the early items correct, the corresponding point-biserial correlations for those items must be near zero. It follows that the overall correlation between WMC and Raven's can only express itself in the point-biserial correlations for the later items where performance is more variable across individuals. Consequently, a greater overall correlation is preferentially observed in the later items, which necessarily translates into an increasing itemwise slope across the entire test.

This increasing slope fails to be observed only if performance on the final test items falls sufficiently close to the floor to constrain their variance, thereby curtailing the itemwise correlations for the last items. The shorter test duration used by Unsworth and Engle (2005) led to near-floor performance on the later test items, thereby preventing the detection of an increasing itemwise slope. This is likely to have been the case even if the overall correlation had been greater. For the increasing slope to be observed, performance on the later items ought to be off the floor and the overall correlation must be large. The standard 40 minute allocation in our study should act to increase accuracy for the later items, and the use of a battery of WM tests should serve to increase the overall correlation between WMC and Raven's performance.

Method

Participants The participants were 130 volunteers (95 female; mean age 21.12) from the University of Western Australia campus community. Participants received either partial course credit for an undergraduate psychology course or $20 for two 1-hour sessions.

Procedure In the first session of the study, participants completed a battery of four WMC tasks (see Lewandowsky, Oberauer, Yang, & Ecker, 2010).

Memory updating (MU). The MU task required participants to (a) store a series of numbers in memory, (b) mentally update these numbers based on a series of arithmetic operations, and (c) recall the updated numbers. On each trial, three to five frames containing random digits were presented on the screen. Following memorization, successive arithmetic operations, (e.g., '+4' or '-3') were presented in the frames, one at a time for a random number of steps before final recall was cued. The key dependent variable is the proportion of updated digits recalled correctly.

Operation span (OSPAN) and Sentence span (SS). On each OSPAN trial, a series of arithmetic equations were presented (e.g., 4+3=7), each of which was followed by a consonant for memorization. Participants judged the equation for correctness and recalled the consonants immediately after list presentation in the original order. The SS task was identical to the OSPAN, except that instead of judging correctness of an equation, participants judged the meaningfulness of sentences (cf. Daneman & Carpenter, 1980). For OSPAN and SS, the key dependent variable is the proportion of consonants recalled correctly.

Spatial short-term memory (SSTM). The SSTM task was adapted from Oberauer (1993) and involved memorization of the spatial location of circles presented, one-by-one, in various locations in a 10 × 10 grid. Participants used the mouse to indicate the memorized location of the dots in any order by clicking in the corresponding grid cells. For this task, participants are given a score based on how similar their recalled pattern was to the to-be-memorized pattern (see Lewandowsky et al., 2010).

Fluid intelligence tests (RAPM) In the second session, participants completed Sets I and II of the 1962 Raven's Advanced Progressive Matrices. As recommended by Raven et al. (1998), RAPM Set I was included to familiarize participants with the matrices. Participants had 5 minutes to complete the 12 items in Set I before being given the standard 40 minutes to complete the 36 items in Set II. We only report the results for the last 36 items (Set II).

Results

Data from two participants who failed to complete all tasks were removed from analysis, and data from two

Table 1: Means *M*, standard deviations *SD*, skewness, and kurtosis for the operation span task (OSPAN), sentence span task (SS), spatial short-term memory task (SSTM), memory updating task (MU), and Raven's Advanced Progressive Matrices (RAPM).

Measure	*M*	*SD*	Skewness	Kurtosis
OSPAN	0.71	0.14	-0.99	4.07
SS	0.70	0.15	-0.70	3.30
SSTM	0.84	0.06	-0.14	2.37
MU	0.66	0.18	-0.34	2.48
RAPM	24.47	5.37	-0.34	2.90

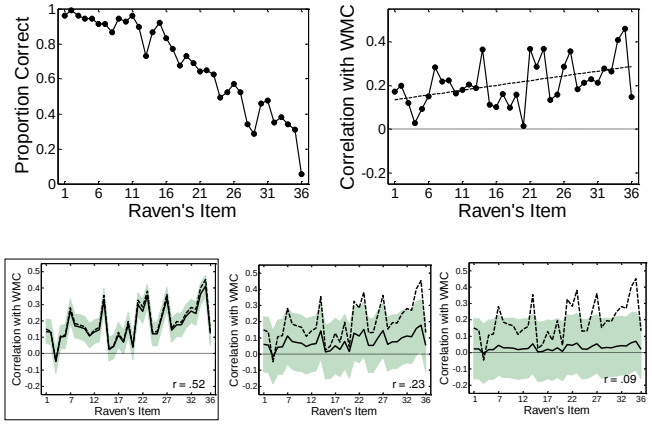

Figure 3: Left: Performance on Raven's Advanced Progressive Matrices items. Right: Observed correlations between working memory capacity (zWMC; based on a battery of four tasks) and performance on each Raven's item. The solid line represents the best-fitting regression line (intercept .13, slope .004). Bottom panels: Results from a bootstrapping analysis resulting in correlations of .52, .23, and .09, respectively, between WMC and overall Raven's performance. All bootstrap results are based on 10,000 replications and the shaded areas represent the 95% confidence regions for the bootstrapped means. The framed bottom-left panel matches the overall correlation and itemwise results in the top right panel.

further participants were discarded for having WMC and Raven's scores less than three standards deviations below the mean, respectively. The final analyses thus used a sample size of $N = 126$. Descriptive statistics for the four WMC tasks and RAPM are shown in Table 1. The top left panel of Figure 3 shows average performance on the RAPM items from Set II. The pattern conformed to expectation in that accuracy decreased with ordinal item position.

WMC and item difficulty For the correlational analyses, we computed a composite measure of WMC by first converting each participant's score on each WM task into a z-score, and then computing that person's average z-score across the four tasks (zWMC). As anticipated, the overall correlation between zWMC and the total RAPM score was moderately large, $r = .56, p < .001$, and larger than was found in previous studies using only a single measure of WMC.

The top right panel of Figure 3 shows the point-biserial correlations between WMC and performance broken down across Raven's items, together with the best-fitting regression line. In contrast to Unsworth and Engle (2005), accuracy was high enough to permit inclusion of all of the Raven's items. The slope of the regression line (.004) was significantly greater than zero, $t(34) = 2.87$, $p < .01$, $r^2 = .20$.[1] The data confirm that when there is at least a moderate correlation between WMC and Raven's performance, the itemwise correlations increase with item difficulty.

Further, to analyse the relationship between zWMC, item difficulty, and performance on Raven's, we conducted a multilevel logistic regression, which circumvents problems due to items with very high or very low accuracy by relying on the logistic (or inverse-logit) function to model the accuracy proportions for each item. We examined a model which includes WMC, ordinal item position (as a proxy for difficulty), and the interaction between these variables.[2] We also systematically tested alternative random-effect models (i.e., letting one or more of intercept or ordinal item position vary randomly across participants.[3]) and determined the preferred model using BIC.

The logistic regression assumes that the predictors are linearly related to the logit transformation of the dependent variable; consequently, we examined the relationship between each variable and accuracy using a White test for nonlinearity (Lee, White, & Granger, 1993). Ordinal item position showed a demonstrable nonlinear relationship with accuracy $\chi^2(2) = 61.12, p < .001$. A Box-Tidwell analysis indicated that the nonlinearity could be removed by raising ordinal item position to a power of 1.704, $\chi^2(2) = 4.29, p = .12$ (see Box & Tidwell, 1962). Because item position is only a proxy for difficulty, the transformation of that variable is acceptable because it retains the ordinal association with the unknown scale of actual difficulty. None of the other variables showed any nonlinear relationship with the largest

[1] The absolute value of this slope is not meaningful because of the relatively large scale of the items (1 to 36) compared to the range of the itemwise correlation.

[2] We could rearrange the items in order of difficulty, but this would bias the analysis towards the result we observe. Instead, we present the items in the order they were presented for parity with Unsworth and Engle (2005).

[3] These random effects models allow the decrease in accuracy across items to begin at a different level of accuracy or, in addition, decrease at a different rate for each participant, respectively

Table 2: Estimated parameters (and standard errors) of mixed effects modeling of the RAPM behavioral study. All significant coefficients are bolded.

Parameters	Model 1	Model 2
Fixed		
Intercept (β_0)	**2.92 (0.11)**	**2.98 (0.11)**
zWMC (β_z)	**0.53 (0.14)**	**0.52 (0.14)**
Item (β_ψ)	**-0.01 (0.0003)**	**-0.01 (0.0004)**
zWMC × Item	**0.001 (0.0005)**	**0.001 (0.0005)**
Random		
Intercept s_0	0.67 (0.59)	-0.05 (0.06)
Item s_ψ		0.0001 (0.0001)
Evaluation		
df	5	7
BIC	4089	4097

χ^2 being for zWMC ($\chi^2\,(2) = 2.86, p = .24$).

Exponentiating ordinal item position to correct for nonlinearity, our first model is given by the following equation:

$$y_{ij} = \beta_0 + \beta_z z_i + \beta_\psi \psi_j^\lambda + \beta_{(z\times\psi)} z_i \times \psi_j^\lambda + (S_i + e_{ij}) \quad (1)$$

where y_{ij} is a binary response variable indicating whether participant i made a correct (1) or incorrect (0) response on item j, z_i is the zWMC score for participant i, ψ_j is the ordinal item position of item j, λ equals 1.704 (as indicated by the above Box-Tidwell analysis), S_i is the set of subject random effects and e_{ij} is an error term.

We tested this model using only the intercept as a random effect (e.g., Model 1, see Table 2) or the intercept plus ψ^λ as random (Model 2). Comparison of the BICs pointed to the model in which only the intercept varied randomly as being preferable (i.e., Model 1). This model revealed significant effects of zWMC ($p < .001$), ordinal item position (ψ^λ, $p < .001$), and the critical zWMC × ordinal item position interaction ($p < .01$). The latter interaction confirms that WMC played an increasingly important role as item difficulty increased, precisely paralleling our initial correlation-slope analysis.

Bootstrapping analysis We next conducted a bootstrapping analysis in which we simulate the effect of decreasing the overall correlation. In other words, to confirm that the magnitude of the overall correlation was responsible for the emergence of an item-difficulty effect in our study, we conducted bootstrapping analyses based on the observed subject × item (126 × 36) response matrix, with rows ordered according to the observed zWMC. The overall correlation between zWMC and Raven's was manipulated by generating new zWMC scores for each participant and examining the effect of that manipulation on the itemwise correlations.

We created three conditions, each involving 10,000 bootstrapping runs. For each run, n, a new vector of zWMC scores was randomly derived from the observed values according to: $z\text{WMC}^{(n)} = \nu\times\ z\text{WMC}+\epsilon$ where $\epsilon \sim N\left(0, \sqrt{(1-\nu^2)}\right)$ and ν varied across conditions. This new vector contained zWMC scores which were derived from the observed zWMC scores but had a reduced correlation with the observed Raven's scores. The rows of the observed binary response matrix were then resorted according to the new vector $z\text{WMC}^{(n)}$ yielding another bootstrapped replication with a specified correlation between zWMC and RAPM that maintained the overall itemwise error rate and overall Raven's correct for each participant observed in the study. Item-wise correlations were then computed between the bootstrapped replication and the zWMC scores.

The three bootstrapping analyses used $\nu = .95$, .50, and .20, respectively, which yielded actual correlations zWMC × RAPM of .53, .23 and .09 (left, center, and right panel in bottom row of Figure 3, respectively). These actual correlations span a large range of possible overall correlations between WMC and Raven's. The bottom left panel provides an idea of the variability expected when the population correlation is approximately equal to that observed in our study. The remaining two panels show that as the population correlation decreases, so does the slope of the itemwise correlations. The center panel roughly corresponds to the correlation observed by Unsworth and Engle (2005) and confirms that the effect of item-difficulty is sufficiently small under those circumstances to escape detection when statistical power is insufficient.

Operation span and RAPM To provide further empirical confirmation that a reduction in the overall correlation between WMC and RAPM attenuates the itemwise effect, we examined the correlation between the OSPAN subtask and RAPM. For this task, the overall correlation with Raven's was much lower, $r = .36, p < .001$. The slope of the regression line for the point-biserial itemwise correlations was not significantly greater than zero, $t(34) = 1.39$, $p = .17$, $r^2 = .05$. Likewise, a multilevel logistic regression (see Equation 1) in which zWMC was replaced by OSPAN failed to find a significant interaction between OSPAN and exponentiated ordinal item position ($p = .09$). This result replicated the invariant relationship found by Unsworth and Engle (2005), supporting our claim the previously published results were obscured by method-specific variance; that is, with a single task, the correlation includes task-specific variance that hides the true magnitude of the underlying correlation between constructs.

Discussion

There were two principal differences between our methodology and previous research. First, we used a

composite measure of WMC which resulted in a higher overall correlation between WMC and Raven's performance. Second, we extended the test time for RAPM to the recommended duration, which resulted in increased overall accuracy. Our results converge on the conclusion that when there is a moderate to strong overall correlation between WMC and performance on the Raven's test of fluid abilities, then the role of WMC becomes increasingly more important as item difficulty increases. Our results suggest that other studies failed to find an effect of item difficulty because in their cases the overall correlation involving WMC was small in magnitude (e.g., Unsworth & Engle, 2005; Wiley et al., 2011). Moreover, the study by Unsworth and Engle (2005) was subtly biased against finding an itemwise effect because of their use of a shorter, non-standard time period for completion of the RAPM test (30 instead of 40 minutes). This non-standard timing made it more likely that performance on the most difficult items would be near the floor (because most people ran out of time before solving those items), thereby necessitating their removal for lack of variance with an ensuant reduction in the power of the analysis.

On the surface, the present work may appear to be merely a statistical issue, but given the intense theoretical attention and interpretation this issue has received, its resolution has considerable psychological implications. In particular, our research cautions against reliance on a null result which has been a substantial barrier to theorizing in this domain. Previously, any model hoping to account for the relationship between WMC and Raven's also had to explain the invariant relationship across item difficulty. The present result shows that this invariance is of questionable generality. By contrast, although not presented here due to space limitations, we have replicated our result using the Raven's Standard Progressive Matrices in another study. Our results therefore open the door for quantitative models of WMC and Raven's that do not predict this invariance. We now know that any model attempting to explain the relationship between the two tasks has to predict that high WMC will allow you to do particularly well on hard items.

Our results are compatible with theoretical analyses of Raven's performance that appeal to working memory as a repository for rules and intermediate results (e.g., Carpenter et al., 1990). Although those theoretical views have fallen out of favor, largely due to the apparent absence of a modulating effect of item difficulty on the relation between WMC and Raven's performance, our results suggest that abandoning those approaches may have been premature.

Acknowledgments

This work was supported by ARC Discovery Grant DP120103888 to the first and second author, an Australian Professorial Fellowship and DORA to the second author and a Jean Rogerson postgraduate scholarship to the third author. We thank Charles Hanich for assistance with data collection and Klaus Oberauer for comments on an earlier version of the manuscript. Address correspondence to the first author at daniel.little@unimelb.edu.au.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, *131*, 30-60.

Box, G. E. P., & Tidwell, P. W. (1962). Transformation of independent variables. *Technometrics*, *4*, 531-550.

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*, *97*, 404-431.

Conway, A. R. A., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review*, *8*, 331-335.

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

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on ackerman, beier, and boyle. *Psychological Bulletin*, *131*, 66-71.

Lee, T.-H., White, H., & Granger, C. W. J. (1993). Testing for neglected nonlinearity in time series models. *Journal of Economics*, *56*, 269-290.

Lewandowsky, S., Oberauer, K., Yang, L.-X., & Ecker, U. K. H. (2010). A working memory test battery for matlab. *Behavior Research Methods*, *42*, 571-581.

Oberauer, K. (1993). Die koordination kognitiver operationen: Eine studie ber die beziehung zwischen intelligenz und working memory (the coordination of cognitive operations: A study on the relation of intelligence and working memory). *Zeitschrift fr Psychologie*, *201*, 57-84.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for raven's progressive matrices and vocabulary scales. section 4: The advanced progressive matrices*. Oxford, UK: Oxford University Press.

Unsworth, N., & Engle, R. W. (2005). Working memory capacity and fluid abilities: Examining the correlation between operation span and raven. *Intelligence*, *33*, 67-81.

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.