1-1-2017

What are standardized literacy and numeracy tests testing? Evidence of the domain-general contributions to students' standardized educational test performance

Steven J. Howard
University of Wollongong, stevenh@uow.edu.au

Stuart Woodcock
Macquarie University, stuart.woodcock@mq.edu.au

John F. Ehrich
Monash University, jehrich@uow.edu.au

Sahar Bokosmaty
University of Wollongong, saharb@uow.edu.au

What are standardized literacy and numeracy tests testing? Evidence of the domain-general contributions to students' standardized educational test performance

Abstract

Background: A fundamental aim of standardized educational assessment is to achieve reliable discrimination between students differing in the knowledge, skills and abilities assessed. However, questions of the purity with which these tests index students' genuine abilities have arisen. Specifically, literacy and numeracy assessments may also engage unintentionally assessed capacities. Aims: The current study investigated the extent to which domain-general processes - working memory (WM) and non-verbal reasoning - contribute to students' standardized test performance and the pathway(s) through which they exert this influence. Sample Participants were 91 Grade 2 students recruited from five regional and metropolitan primary schools in Australia. Methods: Participants completed measures of WM and non-verbal reasoning, as well as literacy and numeracy subtests of a national standardized educational assessment. Results: Path analysis of Rasch-derived ability estimates and residuals with domain-general cognitive abilities indicated: (1) a consistent indirect pathway from WM to literacy and numeracy ability, through non-verbal reasoning; (2) direct paths from phonological WM and literacy ability to numeracy ability estimates; and (3) a direct path from WM to spelling test residuals. Conclusions: Results suggest that the constitution of this nationwide standardized assessment confounded non-targeted abilities with those that were the target of assessment. This appears to extend beyond the effect of WM on learning more generally, to the demands of different assessment types and methods. This has implications for students' abilities to demonstrate genuine competency in assessed areas and the educational supports and provisions they are provided on the basis of these results.

Disciplines

Education | Social and Behavioral Sciences

Publication Details

Howard, S. J., Woodcock, S., Ehrich, J. & Bokosmaty, S. (2017). What are standardized literacy and numeracy tests testing? Evidence of the domain-general contributions to students' standardized educational test performance. British Journal of Educational Psychology, 87 (1), 108-122.

What are Standardised Literacy and Numeracy Tests Testing? Evidence of the Domain-General Contributions to Students' Standardised Educational Test Performance

Howard, S. J.,[1] Woodcock, S.,[2] Ehrich, J.,[3] & Bokosmaty, S.[1]

[1]School of Education, University of Wollongong, New South Wales, 2522, Australia

[2]Faculty of Human Sciences, Macquarie University, New South Wales, 2109, Australia

[3]Faculty of Education, Monash University, Victoria, 3800, Australia

Email addresses: stevenh@uow.edu.au, stuart.woodcock@mq.edu.au, john.ehrich@monash.edu.au, saharb@uow.edu.au

Corresponding author: Steven Howard, School of Education, University of Wollongong, New South Wales, 2522, Australia; Email stevenh@uow.edu.au; Phone +61 2 4221 5165.

Acknowledgements: This research was funded by Internal Funding from the University of Wollongong and Monash University.

Abstract

Background. A fundamental aim of standardised educational assessment is to achieve reliable discrimination between students differing in the knowledge, skills and abilities assessed. However, questions of the purity with which these tests index students' genuine abilities have arisen. Specifically, literacy and numeracy assessments may also engage unintentionally assessed capacities. **Aims.** The current study investigated the extent to which domain-general processes–working memory and non-verbal reasoning–contribute to students' standardised test performance and the pathway(s) through which they exert this influence. **Sample.** Participants were 91 Grade 2 students recruited from five regional and metropolitan primary schools in Australia. **Methods.** Participants completed measures of working memory and non-verbal reasoning, as well as literacy and numeracy sub-tests of a national standardised educational assessment. **Results.** Path analysis of Rasch-derived ability estimates and residuals with domain-general cognitive abilities indicated: (1) a consistent indirect pathway from working memory to literacy and numeracy ability, through non-verbal reasoning; (2) direct paths from phonological working memory and literacy ability to numeracy ability estimates; and (3) a direct path from working memory to spelling test residuals. **Conclusions.** Results suggest that the constitution of this nation-wide standardised assessment confounded non-targeted abilities with those that were the target of assessment. This appears to extend beyond the effect of working memory on learning more generally, to the demands of different assessment types and methods. This has implications for students' abilities to demonstrate genuine competency in assessed areas and the educational supports and provisions they are provided on the basis of these results.

Keywords: standardised test, educational assessment, working memory, reasoning, literacy, numeracy

What are Standardised Literacy and Numeracy Tests Testing? Evidence of the Domain-General Contributions to Students' Standardised Educational Test Performance

Standardised educational assessments are measurement instruments designed to quantify test-takers' abilities in areas such as literacy and numeracy. Their fundamental aim is to achieve reliable discrimination between students who differ in the degree to which they possess the knowledge, skills and abilities assessed. With this information, educational policymakers can determine the extent to which prescribed educational standards are being met and can provide supports to districts, schools and individuals to better meet these standards. As a result, such tests have increasingly been used to make significant decisions related to students, teachers, administrations, communities and schools (Madaus, 1988).

These programs of testing, however, are not without contention. For instance, despite their strong psychometric properties, questions of the purity with which standardised educational assessments genuinely index students' abilities have arisen (Howard et al., 2015; Vilenius-Tuohimaa, Aunola, & Nurmi, 2008; Vista, 2012; Willet & Gardiner, 2009). Specifically, it has been argued that, in the attempt to assess domain-specific literacy and numeracy abilities, standardised assessments may also assess domain-general cognitive capacities (e.g., working memory, non-verbal reasoning; Howard et al., 2015; Willet & Gardiner, 2009). In such cases, it is unclear whether students' results reflect their genuine literacy and numeracy competencies or whether their test scores are constrained by limits in their domain-general cognitive resources (e.g., the approach, method, or phrasing of items overloading students' working memory capacity; e.g., Howard et al., 2015). For example, cognitive load researchers highlight how the complexity of information and its method of presentation can overwhelm children's limited working memory capacity (e.g., Sweller, 2016; Sweller, van Merrienboer, & Paas, 1998). Thus, it may follow that greater domain-

general cognitive demands associated with testing may limit students' ability to demonstrate their emerging academic competencies.

This potential test impurity problem, and by extension the uncertain accuracy of students' test scores, may not initially appear to be a pressing issue given that test results, rankings and reactions nearly always derive from within an assessment. That is, systematic bias within a test may skew students' individual results, but as long as it introduces this measurement error consistently across students it should not impede relative comparisons between years, regions, schools, or students. That said, to identify student needs and provide directed educational support on the basis of standardised assessment results – a fundamental aim of educational assessment – requires that test-takers' strengths and difficulties be truly captured. If students' test scores are instead constrained by non-targeted abilities, educational supports and pathways can be misdirected. As a simple case, a student who fails a numeracy test because of insufficient literacy abilities – rather than a lack of numeracy knowledge or skill – is unlikely to benefit from numeracy remediation. Similarly, if non-verbal reasoning or working memory limit assessment performance, this may not necessarily denote the absence of requisite literacy or numeracy knowledge, or that remediation in this content area will yield assessment gains. There is at least preliminary support for this possibility from research showing that: working memory training can improve numeracy abilities amongst those with low numeracy levels (Kroesbergen, van't Noordende, & Kolkman, 2014); and children with numeracy-related disabilities who do not improve with remediation tend to show immature memory and reasoning strategies (Geary, 1990).

Working Memory, Non-Verbal Reasoning and Standardised Test Performance

One of the most commonly implicated domain-general cognitive abilities in educational assessments is working memory (WM). WM is involved in holding and working with information in mind (Diamond, 2016), which some theoretical accounts separate into verbal

(i.e., phonological) and visual-spatial WM systems (Baddeley, 2002; Jonides, 2000). WM is strongly implicated in acquisition and demonstration of academic abilities, which includes academic achievement (Alloway & Alloway, 2010; Blair, & Razza, 2007), literacy tasks (e.g., Baddeley, 2002; Plaza & Cohen, 2007) and numeracy tasks (e.g., Fuchs et al., 2006; Hecht, Torgesen, Wagner, & Rashotte, 2001; Reuhkala, 2001). In fact, Gathercole and colleagues (2003) found the relationship between primary school students' WM and their performance on Scholastic Aptitude Tests (involving Science, English and Mathematics) was consistently high across all domains (with correlations ranging from .36 to .53). It is therefore expected that WM should be associated with test performance, given its role in learning (e.g., higher WM capacity is associated with more effective, efficient, or complex learning; Bull, Espy, & Wiebe, 2008). Such a relationship would not represent an issue of test impurity, but rather an artefact of the academic achievement gap common between those with high versus low WM (Alloway & Alloway, 2010).

The relationship between non-verbal reasoning and academic achievement is also well established (Gagné & St Père, 2001; Jensen, 1998; Kaufman, Reynolds, Liu, Kaufman, & McGrew, 2012; Rohde & Thompson, 2007). Research indicates that correlations between reasoning and school achievement range from .50 to .70 (Jensen, 1998). In investigating the predictive power of various traits, abilities and demographics, Gagne and St Père (2001) found that reasoning was the strongest predictor of student academic achievement, eclipsing even motivation. Similarly, in an Australian context, Carmichael, Macdonald and McFarland-Piazza (2013) found non-verbal reasoning to be a strong predictor of national standardised test performance, explaining 24.2% of the variability in performance.

Less clear, and more problematic, is the extent to which WM and non-verbal reasoning might also be more directly involved as a result of the design, method, or phrasing of the test itself. This possibility is suggested by studies indicating that the ability to demonstrate one's

knowledge and skills varies as a function of the type of assessment (e.g., Howard et al., 2015; Willet & Gardiner, 2009). For example, Willet and Gardiner (2009) found that 75% of their student participants were better able to spell dictated words than were able to correct visually presented misspelled words (the latter based on Australia's method of assessing spelling in its national standardised assessment). In addition, Howard et al. (2015) used neuropsychological methods to demonstrate experimentally that different modes of assessment can differentially engage WM-related neural networks. Rather than exerting an indirect effect via learning, this renders WM and non-verbal reasoning an unintentional component of, and direct contributor to, students' test performance. It becomes something the tests assess. Instead, it is preferable to maximize variance associated with the underlying abilities of interest, as is the case for all measure and test construction. While it is not possible or preferable to create an educational assessment that does not rely on WM or reasoning at all, unintended measurement error is introduced when domain-general cognitive abilities exert a direct effect on test performance beyond that required for demonstrating the target competency. Even more, the unintentional direct assessment of these non-targeted abilities would not be expected to exert an influence uniformly across the range of test-takers, but instead would enable or constrain performance (independent of content-related abilities being assessed) along a gradient of students' WM capacities.

Estimating Domain-General Contributions to Domain-Specific Assessments

One opportunity for estimating domain-general cognitive contributions to standardised educational assessments derives from the modern test theory approach of Rasch modelling (Andrich, 2004). The Rasch model is commonly used in the social and medical sciences for purposes of scale and test construction. The Rasch model transforms raw data into two independent parameters known as (1) ability estimates and (2) item difficulty estimates. The relationship between the two parameters is expressed on a linear scale as a logistic function

of the relative distance between a person's ability and the difficulty of the item (Pallant & Tennant, 2007). The fit of the data to the Rasch model is determined by proximity of raw score data to the theoretically derived estimates (Pallant & Tennant, 2007).

In contrast, differences between the raw score or observed data and the theoretically derived Rasch estimates are known as residuals. These residuals reflect performance variance that cannot be explained by the construct being measured (such as literacy and numeracy), and thus represent dimensions/constructs outside the unidimensional construct being measured (Tennant & Conaghan, 2007). In a typical Rasch analysis, the standardised person-item residuals are used to determine the unidimensionality of a test or scale. Usually, principal components analysis is conducted on these residuals to identify meaningful patterns in the data (Tennant & Conaghan, 2007). Identification of meaningful patterns in the standardised person-item residuals, in this context, indicates the presence of multiple dimensions being captured (e.g., discrete literacy abilities rather than a singular literacy construct and/or some other unintentionally measured abilities).

The Current Study

Emulating this approach, we hypothesised that Rasch-derived standardised person-item ability estimates and residuals attained from test-takers' standardised test performance could provide a method to investigate the extent to which domain-general processes contribute to standardised test performance, and the pathway(s) through which they exert this influence. Specifically, to evaluate this possibility, Australia's National Assessment Program – Literacy and Numeracy (NAPLAN) was administered to early primary school students, along with a battery of domain-general cognitive tasks. NAPLAN is Australia's nation-wide standardised educational assessment, assessing reading, writing, language abilities and numeracy. Given the substantial research base establishing the role of domain-general processes for content-based learning and achievement, it was expected that WM would have indirect (i.e., through

general ability) and direct associations with NAPLAN-derived person ability estimates. That is, it was expected that: (1) WM would have a significant indirect association with domain-specific ability estimates through non-verbal reasoning; as well as (2) direct associations with Rasch-derived literacy and numeracy ability estimates. Similar results were also expected for Rasch residuals, such that we expected WM would have significant direct and indirect paths to Rasch-derived test residuals (i.e., the cognitive load of assessment), beyond its effects on academic ability. If, as expected, working memory and/or reasoning are strong predictors of NAPLAN standardised person-item residuals, it can be argued that domain-general abilities also support performance on this test, *independent* of the literacy and numeracy constructs being assessed.

Methods

Participants

Participants were 91 Grade 2 students (7-8 years of age), recruited from five regional (n = 3 schools, 29 children) and metropolitan primary schools (n = 2 schools, 63 children) in Australia. The sample was comprised of similar numbers of boys and girls (54.4% female). Participants were all native speakers of English. Parents of participants provided written informed consent as a requirement for participation.

Measures

Standardised Educational Assessment. The standardised assessment administered was Australia's National Assessment Program – Literacy and Numeracy (NAPLAN) test, which is carried out nation-wide with all students in grades 3, 5, 7 and 9 (ACARA, 2011). First established in 2008, NAPLAN assesses students' numeracy, reading, writing and language conventions abilities (i.e., spelling, grammar, punctuation). Annually, students sit this test across multiple days, in a group-testing situation. The 2011 NAPLAN test was adopted and administered because it was the most recent version of this assessment that had no test items

currently in circulation. Specifically, the language conventions (50 written response and multiple choice questions) and numeracy subtests (35 written response and multiple choice questions) were adopted to ensure consistency in marking with national NAPLAN testing.

Domain-General Cognitive Tasks. The *Mr. Ant* visual-spatial working memory task from the Early Years Toolbox (EYT; Howard & Melhuish, 2016) asks children to remember the spatial locations of "stickers" placed on a cartoon ant and then recall these locations after a retention interval. For this task, three trials at each increasing level of difficulty (i.e., Level 1 with a single sticker to Level 8 with eight stickers) progress as follows: (a) Mr. Ant presented with *n* colored stickers (where *n* equals the current level) for 5 s; (b) presentation of a blank screen for 4 s; and (c) an image of Mr. Ant without the stickers presented until the child's response is complete. That the task is administered via iPad allows for the standardisation of timing and scoring. Participants indicate a response by tapping the spatial locations that they think previously held stickers. The task continues until the earlier of completion (at Level 8) or failure on all three trials at the same level of difficulty. WM capacity was indexed by a point score in which: one point is awarded for each consecutive level in which at least two of the three trials were performed accurately; plus 1/3 of a point for all correct trials thereafter.

The *Not This* phonological working memory task from the Early Years Toolbox (EYT; Howard & Melhuish, 2016) asks participants to remember, and then point to, a stimulus that is not of a particular color, shape, or size (or some combination of these). The task consists of five trials at each level of difficulty (from Level 1 with one stimulus feature to hold in mind to Level 8 with eight stimulus features to hold in mind). Each trial proceeds as follows: (a) an auditory instruction played against a white screen (e.g., "Find a shape that is not red"), (b) a 3 s delay against a white screen, and then (c) a 4 × 5 array of differently coloured and sized shapes presented until a response is made by tapping the shape(s) that the participant thinks fulfil the auditory instruction. Directions referring to multiple stimuli must be carried out in

the specified order. The task continues until the earlier of completion (at Level 8) or failure to accurately complete at least three of the five trials within the same level. A point score again indexed performance, calculated as: one point for each consecutive level in which at least three of the five trials were performed accurately; plus 1/5 of a point for all correct trials thereafter.

Raven's Standard Progressive Matrices (RSPM) is widely considered as a measure of non-verbal reasoning and problem solving (Raven, 1976) that is appropriate for both children and adults (age range: 6-80 years). In RPSM, participants are presented with abstract patterns with a single segment that has been removed. Participants must identify the missing part of the pattern from the provided multiple-choice options. The test consists of 60 items (five sets of 12 problems), for which unlimited time is given for completion. RSPM has the additional benefits of being widely used (O'Leary, Rusch, & Guastello, 1991), purportedly more culture fair (Jensen, 1980) and its establishment as a reliable measure of general ability ('*g*'; O'Leary et al., 1991; Raven, Court, & Raven, 1983).

Procedure

All assessments were conducted in quiet rooms of the child's school that were open to visual supervision by school staff. Cognitive tasks were administered in a single, individual testing session, in the following fixed random order to all participants: Raven's, Mr Ant, Not This. Adopting a fixed random task order is consistent with common practice in the cognitive literature, and is the current practice in NAPLAN administration. It was also preferable in the current study given limited statistical power to control for counter-balanced order. NAPLAN was administered in two group sessions per school, one section per day, as follows: language conventions, numeracy (paralleling the order in which NAPLAN is administered; BOSTES NSW 2015). For most participants these sessions were completed on two days in the same week; however, students who were absent for NAPLAN testing completed the missed test on

the next available day upon their return. Performance-derived person-ability estimates and residuals for each score on NAPLAN were provided by NAPLAN's governing organisation, ACARA, for inclusion in statistical modelling.

Results

Initial Data Screening

Initial data screening indicated that 31 participants had at least one missing data point due to absence or early withdrawal. Because exclusion of these participants would have resulted in a loss of 34% of the data, compared to only 8% of the data points being missing, maximum likelihood estimation (i.e., SPSS' expectation-maximisation algorithm) was used to impute missing data. Imputation did not alter the overall pattern of findings, but allowed for generation of fit (i.e., SRMR) and modification indices that require a complete dataset. Confirmatory factor analysis (CFA) using AMOS' maximum likelihood estimation was then used to evaluate the absolute and relative fit of the a priori specified models. In accordance with Hu and Bentler (1998), absolute fit was determined by chi-square statistics and relative fit was assessed using Bentler's comparative fit index (CFI; values > .90 indicating good fit; Smith & McMillan, 2001), standardised root-mean-square residual (SRMR; values < .08 indicating good fit; Hu & Bentler, 1998), root mean square error of approximation (RMSEA; values < .05 indicating good fit; Browne & Cudeck, 1993) and Akaike's information criterion (AIC; lower values indicating comparatively better model fit).

Path Analysis

Spelling. First examined was a spelling model reflecting the hypothesised direct and indirect associations between WM and Rasch-derived ability estimates and residuals (Figure 1). However, poor model fit (Table 1) and the prevalence of non-significant pathways (Table 2) suggested that this model provided poor overall fit to the data. Modification indices for Model 1 suggested an important relationship of ability estimates predicting residuals had not

been modeled. A subsequent model thus added a direct pathway from ability estimates to residuals (Figure 1). While this improved comparative model fit, it still did not yield good overall fit to the data. A third model thus removed the non-significant paths (Figure 1), the result of which provided good fit to the data. In the final model, both visual-spatial and phonological WM were indirectly related to ability levels, through general ability. Visual-spatial WM and ability estimates both also contributed to prediction of the residuals. Given that all paths displayed standardised regression weights of at least .20, and further provided good overall explanatory value of the modeled variables (evidenced by all outcome variables having an $R^2 \leq .15$), this model was adopted as providing good absolute and relative fit to the data.

Grammar. The same a priori initial model of direct and indirect associations between WM, ability estimates and residuals was again evaluated as the initial model for the Grammar data (Figure 1). As with the Spelling model, this model provided poor absolute and relative fit to the data (Table 1). Further, numerous non-significant pathways (Table 2) suggested that this model was insufficient to characterise the data. Modification indices again suggested a relationship of ability estimates predicting residuals, which was subsequently modeled in Model 2 (Figure 1). While this improved model fit, the model remained insufficient to characterise the data. A third model removed all non-significant paths (Figure 1), which provided moderate fit to the data (i.e., CFI, SRMR and AIC indicating good relative fit, whereas RMSEA exceeded levels indicative of 'good' model fit). Inconsistency across the fit statistics further complicated final model selection. Given that a chi-square difference test indicated no significant difference in model fit, $X^2(5) = 10.41$, $p > .05$, the more parsimonious final model was adopted. In this final model, the indirect effects of WM on grammar ability via general ability were maintained (all variables with $R^2 > .25$, suggesting good explanatory value). In contrast, the only predictor of NAPLAN grammar residuals was Rasch-derived

ability estimates (with residuals displaying an $R^2 = .12$, suggesting the limited explanatory value of this model for explaining these residuals).

Numeracy. This same modeling sequence was implemented for the NAPLAN Numeracy data. The a priori model (Figure 1) again provided poor model fit (Table 1) and frequent non-significant pathways (Table 2). A second model with a direct pathway from ability estimates to residuals (Figure 1) improved model fit, but still did not provide good fit to the data. A third model removing the non-significant paths (Figure 1) provided enhanced model fit (albeit falling marginally short of 'good' model fit on RMSEA and SRMR indices). In parallel to the first two models, the indirect pathway from WM to numeracy ability estimates persisted. In contrast to the earlier models, however, phonological WM provided additional direct predictive value for ability estimates. As in the Grammar final model, only Rasch ability estimates predicted residuals. All paths in this final model showed standardised regression weights of at least .19, and further provided good overall explanatory value (indicated by all variables having an $R^2 \leq .25$), thus suggesting good fit to the data.

As a final step, to evaluate the potential for additional literacy demands associated with numeracy test performance (as suggested by the relation of phonological working memory), spelling and grammar ability estimates were additionally modeled as also contributing to numeracy ability estimates (Figure 2). While this model did not provide particularly good overall fit to the data – χ^2 (13) = 39.93, $p < .001$, CFI = .893, RMSEA = .152, SRMR = .159 – these model fit statistics should be considered in the context of the inclusion of additional pathways without an accompanying increase in sample size. Rather, when considering size and significance of the model's pathways, there was clear evidence of spelling and grammar ability contributing to numeracy ability estimates (Figure 2), supporting the suggestion of literacy's role in the numeracy assessment.

Discussion

The current study aimed to determine the direct and indirect contributions of domain-general cognition to performance on Australia's NAPLAN standardised assessment. Results supported the well-replicated relationship between WM and academic performance; however, our results suggest this association was mediated by non-verbal reasoning. These associations were consistent across NAPLAN's spelling, grammar and numeracy assessments. Further, our results also suggested a more-direct influence of phonological WM on numeracy ability and visual-spatial WM on spelling residuals (over and above that explained by ability levels). Subsequent modeling further showed the influence of literacy abilities on numeracy. Taken together, these results suggest that the students' NAPLAN performance was a product of not only the core competencies being assessed, but likely also of a range of non-targeted abilities. This has important implications for whether appropriate educational pathways and supports can be accurately and consistently provided to students on the basis of these assessment results.

A particularly robust finding in the current study was a replication of the well-established effect of WM on learning (Baddeley, 2002) and, by extension, academic performance. In contrast to much of this previous work, however, our results suggest that this effect was mediated by non-verbal reasoning across all of the assessed content domains. Phonological WM also provided additional, direct predictive value for explaining numeracy ability. The fact that phonological WM predicted ability levels only for numeracy might be due to the additional literacy demands of the NAPLAN numeracy assessment. That is, NAPLAN's numeracy questions all questions involve, to some extent, a requirement to unpack a word-based question in order to determine the numeracy concepts that must be applied.

In fact, subsequent modeling to explain this relationship indicated the predictive strength of spelling and grammar abilities on numeracy ability, over and above contributions of WM and reasoning. The association between phonological WM and numeracy ability estimates

thus may be related to the section's literacy demands – a result that is consistent with findings of a predictive relationship between phonological WM and the retrieval, representation and execution of mathematical knowledge and procedures (Geary, 1993), and computation skills (Hecht et al., 2001). While other studies have been unable to replicate this relationship (e.g., Passolunghi, Vercelloni, & Schadee, 2007), these studies often looked at numerical problems directly rather than numeracy more broadly (involving a wider range of numerical concepts, often in the context of worded questions).

As this association with phonological WM was not evident for the spelling or grammar sections, its presence in the numeracy model may be due to the unique literacy demands of this approach to numeracy assessment (that were not present in the literacy assessments). For instance, amongst literacy sub-tests, the spelling test required correction of a misspelt word in which sentence comprehension had little bearing (e.g., there were no homonyms). In contrast, numeracy questions often required decoding and comprehension of complex word problems to extract numerical information prior to being able to generating a response. Similarly, while grammar questions might be expected to impose the greatest literacy demands (QCA, 1998), the fact that these questions all involved multiple choice responses meant that performance required accurate identification of a grammatically correct sentence (instead of generating or copyediting a grammatically correct sentence). As such, it could be argued that the numeracy assessment imposed some of the highest literacy demands, resulting in specific involvement of phonological working memory that could not be accounted for strictly through non-verbal reasoning ability.

Suggestion of the importance of question type was further evident in the association of visual-spatial WM with spelling residuals, but not grammar residuals. That is, visual-spatial WM predicted spelling residual variance that could not be accounted for by spelling ability. In line with previous research (Howard et al., 2015) showing WM is differentially engaged

across test question types, even when holding question difficulty constant, this finding may also be a product of the question type mediating its complexity. Specifically, research has found that WM is more highly engaged during proofreading-type spelling questions relative to spelling a word that is verbally provided (Howard et al., 2015). While that research was limited to spelling assessment, it is notable that NAPLAN's grammar section entailed use of only a multiple-choice format, while the spelling section involved proofreading-type spelling questions. As such, the unique and direct involvement of visual-spatial WM for NAPLAN's spelling performance may be related to additional complexities introduced in proofreading (e.g., identification of a misspelled word, overcoming the provided 'cognitive set' to correct the misspelled word).

This association between WM and spelling residuals further suggests that NAPLAN's proofreading requirement may be peripheral to measuring students' spelling ability. In fact, similar results have been found in the context of reading assessment, which led Rauch and Hartig (2010) to conclude that, given typical practices of attempting to assess higher-level processes through open response formats (in contrast to adopting a closed-choice format to assess basic precursor skills), "response format and assessed skills and cognitive processes are very likely to be confounded in applied assessment" (p. 370). To be clear, the aim is not to eliminate WM demands in standardised educational assessment (or explain 100% of the variance in test performance), nor would it be possible to do so. Rather, indirect effects of WM on academic performance are expected, given the essential role of WM in learning. So too are direct associations between WM and student academic performance expected, when ability estimates are not statistically accounted for. In such cases, WM is a central component of the target abilities (e.g., proofreading requires that students concurrently consider intent, meaning, language conventions, etc.) and test items engage WM in a consistent manner. To the extent that WM explains test residuals, however, these domain-general cognitive abilities

have become a dimension that was directly assessed, yet is peripheral to the target abilities. An important area for further research in educational assessment, beyond typical psychometric evaluation, is therefore to evaluate the extent to which different modes of assessment engage WM in a manner that is essential for the underlying ability of interest and is thus engaged consistently across items.

Given that the current and previous studies are often limited in sample size and by their commonly correlational nature, further research involving experimental research can provide additional clarity around the effect of question type. Specifically, experimental manipulations of existing assessments could provide opportunities for examination of: (1) the acute effects of question type on performance amongst questions that differ only in their literacy demand and response format; (2) latent variable approaches to determine whether questions load more highly on an ability factor on the basis of the skill being assessed or the question type; and (3) a longitudinal approach identifying which question types better predict real-world academic and life outcomes, across ages and content areas. Given the prevailing assumption that a core purpose of assessment is to index students' point-in-time performance along their academic trajectories, which themselves are important to ensure positive later-life outcomes, this form of experimental research could validate the sorts of question types that predict the outcomes that are of interest to parents, practitioners, and policy-makers (a causal relationship that is often assumed, but rarely evaluated at the design-level of education assessment).

Research should also consider various existing assessments across the schooling years, to evaluate the extent to which these relationships may change with age, learning and familiarity with standardised educational assessment. The current study's administration of the Grade 3 NAPLAN test to Grade 2 students, for instance, may have inflated the degree to which WM was engaged during the students' performance of this test. While this would not be expected to influence the overall pattern of associations, and had the benefit of engaging students naïve

to formal standardised testing, it is important to examine these associations across the school years. This research would also benefit from considering a broader range of factors that may account for students' standardised test performance. While the current study highlights the domain-general cognitive factors that can contribute to performance, there remained a fair degree of unexplained 'error' variance. Factors such as student motivation and test anxiety, for example, may account for at least some of this unexplained error variance (Mavilidi, Hoogerheide, & Pass, 2014; Wolf & Smith, 1995).

The current study provides preliminary evidence of the contributions of WM and non-verbal reasoning to students' NAPLAN performance – Australia's annual and nation-wide standardised educational assessment. Specifically, our results suggest that this method of assessment confounded non-targeted abilities with those that were the target of assessment. This appears to extend beyond indirect associations with ability levels as a result of WM's influence on learning, but also to the demands placed by the assessment type and method. This has implications for students' abilities to demonstrate genuine competency in the assessed areas, with follow-on effects for the educational supports and provisions they are provided on the basis of these results. While the general aims of standardised assessment are laudable, undertaking this program of assessment carries additional responsibility to ensure the accurate capture of students' abilities. If assessments fail in this regard, practitioners may spend time focusing on fostering proficiency in areas that are less likely to influence student outcomes and students may be misconstrued as requiring supports that do not address their true underlying needs. The current study suggests the importance of further work to identify the content areas and forms of assessment that better capture the trajectories of students' core academic competencies.

References

ACARA. (2011). NAPLAN. Retrieved January 31, 2012, from Australian Curriculum Assessment and Reporting Authority (ACARA): http://www.nap.edu.au/naplan/naplan.html.

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology 106*, 20–29. doi: 10.1016/j.jecp.2009.11.003

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1), 7–16. doi: 10.1097/01.mlr.0000103528.48582.7c

Baddeley, A. D. (2002). Is working memory still working? *European Psychologist, 7(2),* 85 – 97. doi: 10.1027//1016-9040.7.2.85

Blair, C., & Razza, R.P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development, 78*(2), 647-663. doi: 10.1111/j.1467-8624.2007.01019.x

Board of Studies Teaching and Educational Standards NSW (BOSTESNSW). (2015). *NAPLAN*. Retrieved from http://www.boardofstudies.nsw.edu.au/naplan/

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, US: Sage.

Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. *Developmental Neuropsychology, 33*(3), 205-228. doi: 10.1080/87565640801982312

Carmichael, C., Macdonald, A., & McFarland-Piazza, L. (2013). Predictors of numeracy performance in national testing programs: Insights from the longitudinal study of Australian children. *British Educational Research Journal.* DOI:10.1002/berj.3104

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*, 13–21. doi: 10.1016/j.intell.2006.02.001

Diamond, A. (2016). Why assessing and improving executive functions early in life is critical. In J. A. Griffin, P. McCardle, & L. S. Freund (Eds.), *Executive function in pre-school age children: Integrating measurement, neurodevelopment, and translational research*. Washington, DC: APA. doi: 10.1037/14797-002

Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., Schatschneider, C., & Fletcher, J. M. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic, computation, and arithmetic word problems. *Journal of Educational Psychology, 98*(9), 29 – 43. doi: 10.1037/0022-0663.98.1.29

Gagné, F., & St Père, F. (2001). When IQ is controlled, does motivation still predict achievement? *Intelligence, 30*, 71–100. doi: 10.1016/S0160-2896(01)00068-X

Gathercole, S. E., Pickering, S. J., Knight, C., & Stegmann, Z. (2003). Working memory skills and educational attainment: Evidence from national curriculum assessments at 7 and 14 years of age. *Applied Cognitive Psychology, 18*, 1-16. doi: 10.1002/acp.934

Geary, D. C. (1990). A componential analysis of an early learning deficit in mathematics. *Journal of Experimental Child Psychology, 49*, 363 – 383. doi: 10.1016/0022-0965(90)90065-G

Geary, D. C. (1993).Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin, 114*, 345 – 362. doi: 10.1037/0033-2909.114.2.345

Guttman, L. (1950). The problem of attitude and opinion measurement. In S. A. Stouffer (Ed.), *Measurement and Prediction*. New York: Wiley.

Hazbic, D., Holzworth, D., & Berry, B. (2014). Cognitive abilities underpin academic performance: Australian secondary school study. *Neuromite Research*. Retrieved from http://www.neuromite.com.au/wp-content/uploads/2015/03/NEUROMITE-Research-Abilities-and-Academic-Achievement.pdf

Hecht, S. A., Torgesen, J. K., Wagner, R. K., & Rashotte, C.A. (2001). The relations between phonological processing abilities and emerging individual differences in mathematical computation skills: A longitudinal study from second to fifth grades. *Journal of Experimental Child Psychology, 79,* 192 – 227. doi: 10.1006/jecp.2000.2586

Howard, S. J., & Melhuish, E. C. (2016). An Early Years Toolbox (EYT) for assessing early executive function, language, self-regulation, and social development: Validity, reliability, and preliminary norms [online first]. *Journal of Psychoeducational Assessment*. doi: 10.1177/0734282916633009

Howard, S. J., Burianova, H., Ehrich, J., Kervin, L., Calleia, A., Barkus, E., Carmody, J., & Humphry, S. (2015). Behavioural and fMRI evidence of the differing cognitive load of domain-specific assessments. *Neuroscience, 297,* 38-46. doi: 10.1016/j.neuroscience.2015.03.047

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Jonides, J. (2000). Mechanisms of verbal working memory revealed by neuroimaging studies. In B. Landau, J. Sabini, J. Jonides and E. Newport (Eds.), *Perception, Cognition, and Language Essays in Honor of Henry and Lila Gleitman.* (pp. 87 – 105). Cambridge, Massachusetts: The MIT Press.

Kaufman, S. B., Reynolds, M. R., Liu, X., Kaufman, A. S., & McGrew, K. S. (2012). Are cognitive g and academic achievement g one and the same g? An exploration on the

Woodcock–Johnson and Kaufman tests. *Intelligence 40*, 123–138. doi: 10.1016/j.intell.2012.01.009

Kroesbergen, E. H., van 't Noordende, J. E., & Kolkman, M. E. (2014). Training working memory in kindergarten children: Effects on working memory and early numeracy. *Child Neuropsychology, 20*(1), 23-37. doi: 10.1080/09297049.2012.736483

Lingard, R. (2010). Policy borrowing, policy learning: testing times in Australian schooling. Critical Studies in Education, 51(2), 129–145. doi: 10.1080/17508481003731026

Loehlin, J. C. (1992). *Latent variable models: An introduction to factor, path, and structural analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education, 65*(3), 29-46. doi: 10.1080/01619568809538611

Mavilidi, M.-F., Hoogerheide, V., & Paas, F. (2014). A quick and easy strategy to reduce test anxiety and enhance test performance. *Applied Cognitive Psychology, 28*(5), 720-726. doi: 10.1002/acp.3058

O'Leary, U, M., Rusch, K. M., & Guastello, S. J. (1991). Estimating age-stratified WAIS-R IQs from scores on the Raven's Standard Progressive Matrices. *Journalof Clinical Psychology, 47*, 277 – 284. doi: 10.1002/1097-4679(199103)47:2<277::AID-JCLP2270470215>3.0.CO;2-I

Pallant, J., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*(1), 1 - 18. doi: 10.1348/014466506X96931

Passolunghi, M. C., Vercelloni, B., & Schadee, H. (2007). The precursors of mathematics: Working memory, phonological ability and numerical competence. *Cognitive Development, 22*(2), 165 – 184. doi: 10.1016/j.cogdev.2006.09.001

Plaza, M., & Cohen, H. (2007). The contribution of phonological awareness and visual attention in early reading and spelling. *Dyslexia, 13*(1), 67 – 76. doi: 10.1002/dys.330

Qualifications and Curriculum Authority (QCA). (1998). *The grammar papers: Perspectives on the teaching of grammar in the national curriculum.* London: Qualifications and Curriculum Authority.

Rauch, D. P. & Hartig, J. (2010). Multiple-choice versus open-ended test format of reading test items: A two dimensional IRT analysis. *Psychological Test and Assessment Modeling (52)*, 354-379.

Raven, J.C. (1976). *Standard progressive matrices*. Oxford: Oxford Psychologists Press.

Raven, J. C., Court, J. H., & Raven, J. (1983). *Manual for Raven's Progressive Matrices, and Vocabulary Scales, Part Three, Section 7, Research.* London: H. K. Lewis.

Reuhkala, M., (2001). Mathematical skills in Ninth-graders: Relationship with visuo-spatial abilities and working memory. Educational Psychology 21(4), 387-399. doi: 10.1080/01443410120090786

Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence, 35*, 83–92. doi: 10.1016/j.intell.2006.05.004

Sweller, J. (2016). Cognitive Load Theory, evolutionary educational psychology, and instructional design. In D. C. Geary & D. B. Berch (Eds.), *Evolutionary perspectives on child development and education*. Cham, Switzerland: Springer. doi: 10.1007/978-3-319-29986-0_12

Sweller, J., van Merrienboer, J. G., & Paas, G. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, (10),* 3, 251 - 296. doi: 10.1023/A:1022193728205

Tennant, A., & Conaghan, P., G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research, 57*(8), 1358 – 1362. doi: 10.1002/art.23108

Vilenius-Tuohimaa, P. M., Aunola, K., & Nurmi, J. E. (2008). The association between mathematical word problems and reading comprehension. *Educational Psychology: An International Journal of Experimental Educational Psychology, 28*(4), 409 – 426. doi: 10.1080/01443410701708228

Vista, A. (2012). *The role of problem-solving ability and reading comprehension skill in predicting growth trajectories of mathematics achievement of ESB and NESB students.* Melbourne: University of Melbourne (PhD thesis).

Willet, L., & Gardiner, A. (2009, July). Testing spelling – exploring NAPLAN. *Paper presented at the Australian Literacy Educators Association Conference*.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*(3), 227-242. doi: 10.1207/s15324818ame0803_3

Table 1

Path Analysis Model Fit Indices

Model	df	χ^2	RMSEA	CFI	AIC	SRMR
Spelling Model						
1. Initial model	2	53.15*	.533	.517	89.15	.154
2. Add Ability -> Residual	1	4.46*	.196	.967	42.46	.066
3. N/S Paths Removed	5	7.15	.069	.980	37.15	.072
Grammar Model						
1. Initial model	2	50.25*	.518	.454	86.25	.161
2. Add Ability -> Residual	1	4.46*	.196	.961	42.46	.063
3. N/S Paths Removed	6	14.87*	.128	.900	42.87	.105
Numeracy Model						
1. Initial model	2	65.48*	.594	.574	101.48	.151
2. Add Ability -> Residual	1	4.46*	.196	.977	42.46	.069
3. N/S Paths Removed	5	8.52	.088	.976	38.52	.093

Note. Model fit is considered as good/comparatively better if: χ^2 is lower and non-significant; RMSEA < .05; CFI > .90; lower AIC; and, SRMR < .08. *p < .05.

Table 2

Factor Loadings for SEM Numeracy Models

	Spelling			Grammar			Numeracy		
Path **Model #**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
VWM -> Reason.	**.30**	**.30**	**.30**	**.30**	**.30**	**.30**	**.30**	**.30**	**.30**
PWM -> Reason.	**.41**	**.41**	**.41**	**.41**	**.41**	**.41**	**.41**	**.41**	**.41**
Reason. -> Ability	**.41**	**.41**	**.43**	**.42**	**.42**	**.35**	**.43**	**.43**	**.48**
Reason. -> Residual	-.22	.05	-	-.10	.18	-	**-.30**	.05	-
VWM -> Ability	-.04	-.04	-	-.15	-.15	-	.17	.17	-
PWM -> Ability	.08	.08	-	-.05	-.05	-	**.19**	**.19**	**.19**
VWM -> Residual	**.24**	**.22**	**.21**	.10	.00	-	-.11	.02	-
PWM -> Residual	-.17	-.12	-	.14	-.10	-	-.12	.03	-
Ability -> Residual	-	**-.67**	**-.68**	-	**-.68**	**-.60**	-	**-.80**	**-.76**

Note. All factor loadings are standardised regression weights. VWM = visual-spatial working memory (Mr. Ant). PWM = phonological working memory (Not This). G = general ability (Raven's). Ability = Rasch-derived person-ability estimates (NAPLAN). Residual = Rasch-derived residuals. (NAPLAN) Significant paths have been bolded.

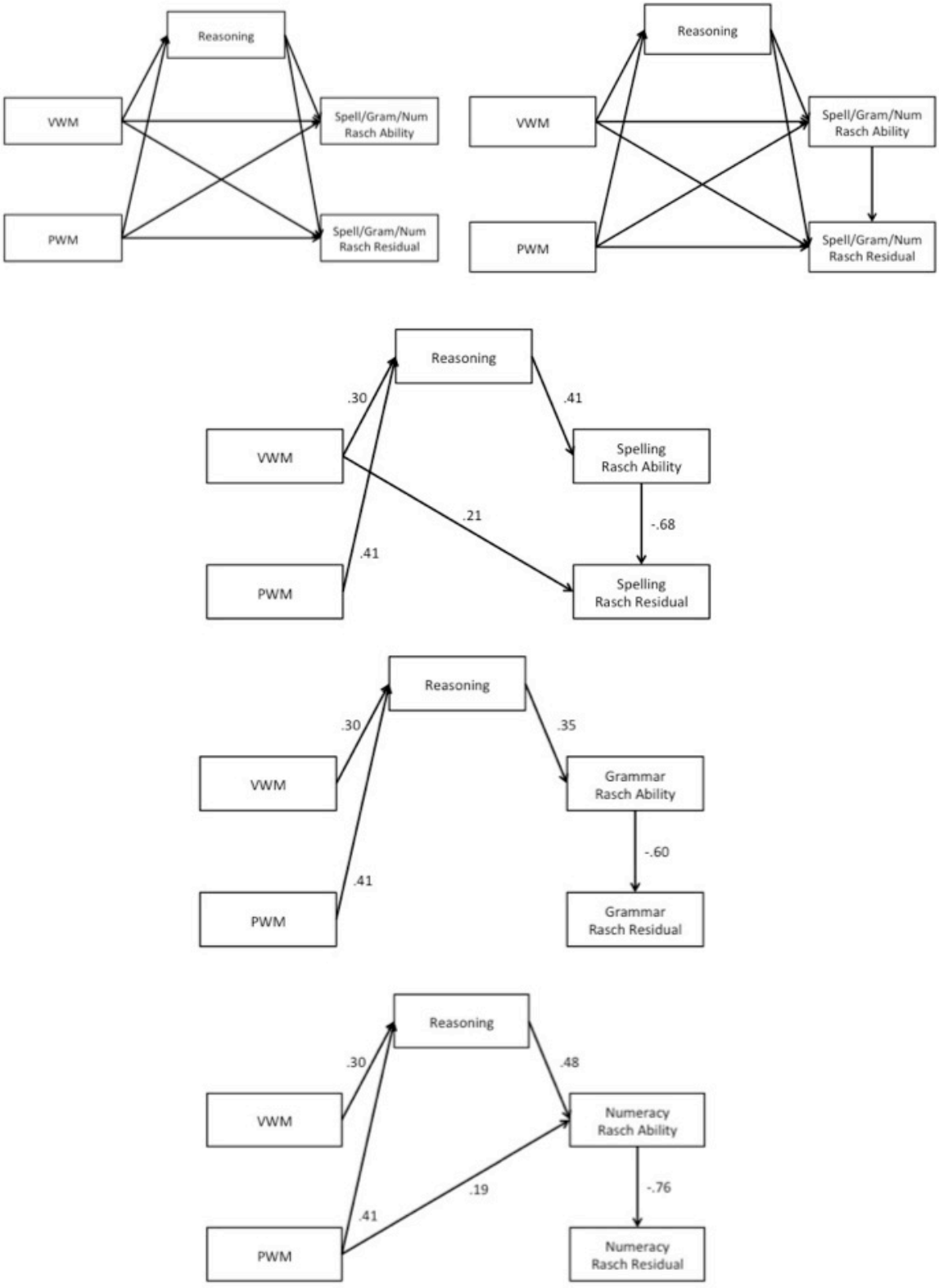

Figure 1. Depicted are Model 1 for all three academic outcomes (top left), Model 2 for all three academic outcomes (top right), and then (in order from top to bottom) Model 3 for spelling, grammar and numeracy. Path loadings are standardised regression weights.

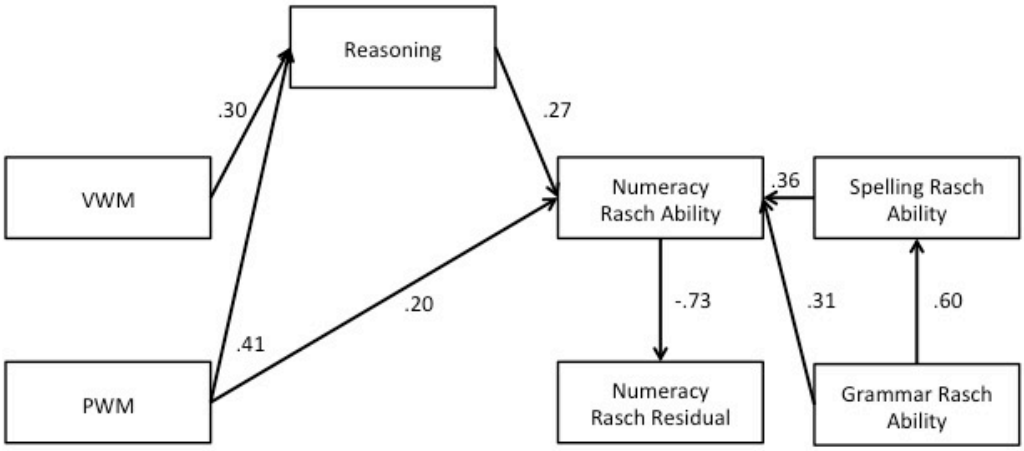

Figure 2. Exploratory numeracy model incorporating spelling and grammar ability estimates. Path loadings are standardised regression weights.