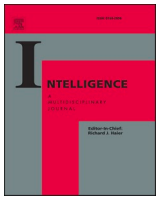
Intelligence 90 (2022) 101604

Contents lists available at ScienceDirect

Intelligence

journal homepage: www.elsevier.com/locate/intell



Process differences as a function of test modifications: Construct validity of Raven's advanced progressive matrices under standard, abbreviated and/or speeded conditions – A meta-analysis

Corey E. Tatel, Zachary R. Tidler, Phillip L. Ackerman *

Georgia Institute of Technology

ARTICLE INFO

Keywords:
Speed
Level
Raven's progressive matrices
Short-form tests
Test modification

ABSTRACT

Historically, there has been substantial disagreement about the importance of speed vs. level in determining individual differences in intelligence – a disagreement that persists across various different modern assessment measures of intellectual abilities. The current investigation considers whether changes to the administration constraints (time limitations or speededness, and total test length) of the Raven's Advanced Progressive Matrices test – which has been identified as a measure highly saturated with general intelligence – results in differences to the underlying ability determinants of test performance. A review of empirical studies was conducted, where versions of Raven's Advanced Progressive Matrices Tests were administered under various time constraints and item lengths. Meta-analytic techniques were used to determine whether introducing speed constraints or shortening the length of the test changes the construct validity of the tests (as indicated by differences in convergent and discriminant correlations with other ability traits). The meta-analysis combined results from 142 studies composed of a total of 26,848 participants. Substantial differences were found for correlations of Raven's Advanced Progressive Matrices and Spatial Visualization (as large as $\widehat{\rho} = 0.26$), Memory (as large as $\widehat{\rho} = 0.08$), and Perceptual Speed (as large as $\widehat{\rho} = 0.34$) abilities under speeded conditions and shorter test lengths. Examinees may draw on different strategies for test performance, that in turn, draw on different combinations of abilities, when the test is abbreviated or significant time constraints are introduced. Implications for using this test under different conditions are discussed.

1. Background

The earliest modern tests of general intelligence (Binet & Simon, 1905/1961) were 'power' tests in the strictest sense. Test items were arranged in difficulty from the easiest items that could be answered correctly by most or all of the examinees, to those sufficiently difficult to be only answered by high-ability children of the same age as the examinee (or older children). An examinee's score on the test was the boundary point between items answered correctly and those (more difficult) items answered incorrectly. In addition to the ordering of items by difficulty, the other main characteristic of power tests is the lack of a time limit for completion. Current intelligence tests modeled on the Binet-Simon Scales are designed to be administered in a one-on-one setting, and they remain power tests with few time limits. However, during the 1910s, with the introduction of "group" tests of intelligence – most notably, the Army Alpha Test (see Yoakum & Yerkes, 1920), strict testing time limits were imposed. Although the items of the Army Alpha Test were largely ordered in terms of difficulty as with power tests, the examinees were instructed to: "...get the answers ... as quickly as you can" (p. 220).

The introduction of such time limits on intelligence tests can be considered from three major perspectives: (1) practical, (2) criterion-related validity, and (c) theoretical (construct validity). Each of these will be treated briefly in turn below.

1.1. Practical aspects of time-limited tests

The practical basis for choosing time-limited vs. unspeeded tests in intelligence assessment is relatively obvious, at least for group-testing scenarios. That is, in a one-on-one intelligence test like the Stanford-Binet or the Wechsler tests, assessment may take a few hours, but the administration proceeds at a pace that is appropriate to the examinee.

* Corresponding author at: School of Psychology, Georgia Institute of Technology, 654 Cherry Street, MC 0170, 30319 Atlanta, GA, USA.
E-mail address: plackerman@gatech.edu (P.L. Ackerman).

https://doi.org/10.1016/j.intell.2021.101604
Received 25 April 2021; Received in revised form 12 September 2021; Accepted 15 November 2021
Available online 21 December 2021

Plus, in a pure power test administration format, the number of test items actually administered to the examinee is typically substantially fewer than would be administered in a group-testing scenario, because each scale is terminated when the examinee can no longer answer more difficult questions correctly. Thus, there is little down-time where an examinee would have nothing to do – the examiner simply moves from one component test to the next.

Traditional group-testing scenarios have constraints that are not encountered in one-on-one testing situations. As noted by Carroll (1982), group tests typically shift their demands on the examinee from 'construction' or 'recall' of correct test answers, to 'recognition' of correct answers from multiple choices, which may tap fundamentally different underlying psychological processes. Also, within a group of examinees, a wide range of talent is ordinarily expected. As such, some lower-ability examinees might take an exceptionally long time in order to attempt all of the test items. In contrast, one can expect a significant number of examinees will complete all items much faster than other examinees, and will be sitting in the examination room with nothing to do while the lower-ability examinees complete the test.

1.1.1. Criterion-related validity

Consideration of criterion-related validity for power vs. time-limited tests is perhaps the most straightforward problem to be addressed, yet one that has received only a moderate amount of attention. There are, however, two elements that need to be considered in order to assess the impact of test format in a criterion-related validity application: (1) a direct comparison between power and time-limited versions of the same test, and (2) an indirect method that considers the differences in total administration time between power and time-limited tests and the consequent opportunity cost/savings. The most basic direct assessment of the comparative criterion-related validity would involve two independent samples, one completing a power version of a test, the other receiving a time-limited test. One could evaluate the advantage by comparing the respective test version correlations with the criterion for both samples. Evaluations could be also made with different testing time limits for multiple samples, in order to determine what the optimal time limit is for maximizing criterion-related validity. More nuanced evaluations could be made by determining the accuracy of predictions made at various cut-off scores.

The indirect method should not be ruled-out in evaluating whether criterion-related validity can be maximized with a time-limited version of the test, rather than a power version of the test. That is, even if the time-limited test has a lower criterion-related validity than the power version, the savings of administration time could be used to include a different measure of the same or other related constructs that also have predictive validity for the criterion. So, an apples-to-apples comparison might consider 'what else' could be included in a selection battery within the total time it would take to administer the power version of the test.

The Wittmann and Süß (1999) conceptualization of Brunswik Symmetry is a useful framework for developing hypotheses about whether power or time-limited versions of a test are likely to have greater criterion-related validity. This perspective specifies that when the predictor and criterion have maximum overlap of the breadth and content of the sources of variance, there will be an optimal level of validity. Presuming that the underlying content of the test is similar between the power and time-limited versions of the test, the major determinant of comparative validity will be the speededness of the criterion. An unspeeded criterion would be expected to show highest validity for the power version of the predictor test, but a criterion that depends on rapid problem solving, for example, would be expected to have higher degrees of shared variance with a time-limited version of the predictor measure, all else being equal.

1.2. Theoretical aspects of time-limited tests

Although early omnibus tests of intelligence, as discussed earlier, did not have notable demands on speed of answering questions, nor do they provide substantial credit for answering questions quickly in contrast to answering correctly, early modern researchers did devote attention to the construct of speed in the context of the construct of intelligence. As E. L. Thorndike et al. (1926) noted:

> "If speed deserves any weight in determining the measures of intellect it is by virtue of the principle that, '*Other things being equal, the more quickly a person produces the correct response, the greater is his intelligence.*'" (p. 24).

However, they went on to note an important limitation of extant tests in that:

> "... a battery of tests in which *level, extent,* and *speed* combine in unknown amounts to produce a test score may be very useful. For rigorous measurements, however, it seems desirable to treat these three factors separately, and to know the exact amount of weight given to each when we combine them." (p. 25).

From this perspective, time-limited tests may have an advantage of producing scores that represent a combination of the level of intelligence and the speed of intellectual processes, when compared to power tests. But, as noted by Thorndike and his colleagues, there is an unknown weighting of level and speed influences on any particular test score, and that would be potentially influenced by changes to the time limits imposed by the test developer.

It should be noted in passing that when considering tests of higher-order mental abilities, the 'speed' considerations are not the same as the constructs described by others that pertain more to Perceptual Speed (PS) or Psychomotor Speed (PM) abilities (e.g., see Ackerman & Beier, 2007; Ackerman, Beier, & Boyle, 2002). In the PS/PM domain, tests are nearly always highly limited in administration time, and the items are generally easy enough to be answered without error if the time limits are relaxed. Relaxing time limits for such tests is expected to result in near-ceiling levels of performance from most examinees.

In the case of complex problem-solving or comprehension tests, introducing time limits may have a negligible effect on examinee performance, perhaps up until the point where insufficient time is allowed for some individuals to even attempt some of the items within the imposed time limit. A vocabulary test might show this kind of effect, given that performance on an unspeeded version yields similar results to performance on the test with moderate time limits (e.g., see Lord, 1956). In other tests, however, an imposition of time-limits may result in changes in how the examinees process test items, especially in conjunction with particular trade-off functions for correct responses and errors (e.g., see Quereshi, 1960). That is, for a test with no penalties for guessing, an examinee might check the amount of time remaining in the test, and then make random responses to any otherwise unattempted items, in an effort to maximize his/her overall test score. If a penalty for guessing is imposed that is greater than the probability of success of a guess (e.g., greater than 0.25 points for a four-choice multiple choice item), then an examinee who seeks to maximize overall test score would not be motivated to make random responses for unattempted items, but rather devote all of his/her attention to at least eliminating some of the unlikely response options before choosing to answer a test item. Ultimately, the imposition of time limits for such tests, along with particular error penalty instructions, may combine to introduce both overall strategy effects and even individual differences in the use of various guessing strategies, depending on the individuals' understanding and ability to implement optimizing approaches for maximizing overall test score.

As noted by Thorndike et al. (1926), in the absence of specific information about the contributions of level and speed sources of variance

in test performance, it is unknown how much each source contributes to overall test performance. It follows that there may be consequential changes to the reliability and validity of the test, should it be shifted from a relatively untimed power test to a time-limited test. If, for example, there is no penalty for guessing, then some individuals who cannot solve all of the problems and respond randomly or nearly randomly to multiple items should result in a reduced internal consistency reliability for the test, given that it would consist of a mixture of items attempted via the use of problem-solving abilities and random responses. Similarly, if the test includes penalties for wrong answers, then the test scores may be influenced by a mixture of those items attempted via problem-solving abilities and those where, rather than solving the problems directly, the examinees have at least sought to eliminate one or more of the response options (which may represent an ability related to the test content, or a general 'method factor').

Resolution of these issues from a construct validity perspective might be accomplished with a multi-method, multi-trait approach, where the methods are different degrees of unspeeded or time-limited test administration formats, and the traits are similar levels of complexity (e.g., if the target test taps spatial reasoning ability, then other related traits might tap numerical and verbal reasoning abilities). Presumably, the different trait measures with the same level of speededness would have identical penalties for guessing and the same or similar levels of environmental press (i.e., the same or similar time-to-solution items). With such a design, it would be possible to determine convergent and discriminant validity of the target test, and how much variance in 'method' is accounted for by the unspeeded or time-limited test types, at least within a single score penalty payoff system.

1.3. Completion time for one intelligence problem

Setting time limits for tests is, however, not *entirely* a matter of partitioning observed test scores into the underlying construct variance (which may or may not include speed as an aspect of construct) and any separate psychological speed components. One prominent concern has to do with the identification of a lowest-level unit of analysis for the construct under consideration. For example, researchers at least since E. L. Thorndike (1908) have considered mental multiplication as an aspect of intellectual ability. In its simplest form, mental multiplication tasks would contain single-digit by single-digit multiplication problems. Solution of these problems can be accomplished by most adults with minimal cognitive mediation, because the single-digit times tables have been memorized to a level of automaticity. In fact, a test of single-digit mental multiplication problems would ordinarily only minimally be considered a test of numerical abilities, but rather a test of PS, because the number of items answered in a fixed period of time is probably limited mostly by the speed of writing down the answers. Because of the limited number of potentially unique problems and the speed with which they can be answered, the only suitable test with such content would be a highly time-limited (speeded) test.

Extending the test to include two-digit x two-digit items represents a fundamental change in how the problems are solved, and the amount of time needed for most individuals to solve them. Thorndike's examinees were presented with even more complex three-digit × three-digit multiplication problems. The average completion time for a *single* problem of this type was about sixsix–seven min. More complex items can also be created and administered, such as Arai's (1912) four-digit × four-digit mental multiplication test, where adult examinees required approximately 10 min or more to complete each item. In fact, just the process of memorizing the problem after presentation prior to starting the solution process likely involves more complex processes for a three or four-digit mental multiplication problem than a one or two-t-wodigit mental multiplication problem.

All of the above examples are fundamentally tests of 'mental multiplication.' But, given that increasing the number of digits to be multiplied in a problem results in demands for fundamentally different cognitive processes or different degrees of integration across processes, selection of one type of test item will inevitably constrain the minimum amount of time required to assess the underlying ability, because a test will require multiple items to achieve acceptable levels of reliability and validity. A test of one-digit mental multiplications might have dozens of items and be administered with a one–two min time limit. But, a test of three-digit × three-digit mental multiplication items would likely require close to an hour minimum to administer 10 or so items.

Other examples include tests of reading comprehension (where comprehension of a single sentence may be fundamentally different from comprehension of an entire paragraph or even a short story), writing, or simple-to-complex problem solving. It is, of course, an empirical question whether there are significant differences in construct validity for tests that vary in terms of the depth and breadth of test item content.[1] Ultimately, the point here is that the choice of any particular time limit for a test will be constrained by the minimum amount of time to complete a single item, and the total number of items needed to achieve acceptable levels of reliability and validity. Setting an arbitrary time limit for a test, in turn, will constrain the depth and breadth of the intellectual demands imposed by the test items. That is, a test that has a 10-min fixed time limit will not be capable of sampling the complex intellectual functions described above.

1.4. The psychological test as an experiment

In the early part of the last century, Terman (1924) surveyed several leading psychologists as to whether the psychological test could be considered as an 'experiment', in the methodology of experimental psychology. E. G. Boring's response (as quoted in Terman) was that "methodologically there is no essential difference between a mental test and a scientific psychological experiment...." (p. 98). As noted by Ackerman and Lohman (2006), "At a narrow level, this refers to the fact that each test item is a stimulus, and each answer is a response, in the classic behaviorist representation." (p. 151). From this perspective, a change in any of the conditions of testing (e.g., instructions, number test items, time limits, ordering of items, paper and pencil vs. computerized administration) represents a potentially significant change in the 'experiment' underlying the test, which in turn, may change the dependence of the test scores for a group of examinees on other abilities or non-ability traits (e.g., personality, interests).

Although there are numerous articles in the literature about the effects of changes in conditions of testing for consequential tests of IQ, such as the Stanford-Binet and the Wechsler tests, some investigators appear to have overlooked many of the considerations from the experimental-psychology perspective, which historically would require a theoretical and empirical review of the effects of changes to an experimental paradigm, and depended instead on considerations of the *efficiency* of abbreviated tests. Such an approach, without careful review of the potential effects of changing testing conditions may have unintended consequences, especially in terms of the constructs underlying test performance.

Ideally, an examination of whether changes in the speededness of test administration results in differential convergent and/or discriminant validity would involve consideration of multiple tests representing a variety of different abilities-to-be-measured. Practically, such an examination is not possible by reviewing the literature in the form of a meta-analysis. The main reason is that there are few ability measures

[1] Freeman (1928) made a similar argument in explaining why the analogy of a foot race is an inappropriate analogy for the effects of different time-limits for tests of intellectual ability. He noted that there is a fundamental difference in the physical abilities required for a race to be run in 10 or 20 s – indeed, it is not unusual to find differences in rank-ordering of runners for such different races, and it is even more typical to find rank-order differences for races of 100 m and 10,000 m, even though the underlying process is to run as fast as one can.

that have been administered in a variety of different formats with sufficient samples and that include assessments of other abilities, so that differences in construct validity can be assessed. There are some large-scale studies concerning the effects of extended time with college selection measures (SAT, ATC, LSAT, GRE; see, for example, Evans & Reilly, 1972; Wild, Durso, & Rubin, 1982), but these investigations have focused mainly on changes in mean scores with extended testing time compared to standard time limits, and not with construct validity of the tests. The only intellectual ability test for which we could find adequate samples of studies with varying degrees of speed limitations and correlations with a variety of different reference ability measures was the Raven's Advanced Progressive Matrices (RAPM) test. The advantage of investigating this test is relatively obvious, because some researchers claim that the test is essentially a 'prototypical' measure of fluid intelligence (*Gf*) (Birney, Beckmann, Beckmann, & Double, 2017; Salthouse, 2014), while others have suggested that the test is less a measure of *Gf*, but includes influences from a variety of other abilities (e.g., spatial, memory, perceptual speed) (Burke, 1958; Gignac, 2015). Thus, we focused on the RAPM as the central source of our investigation of determining whether the conditions of testing are related to differentiation of the abilities underlying performance on the test.

1.5. The current investigation

Because of the extensive literature on the RAPM that includes multiple 'short-form' modifications with a variety of different administrative constraints in terms of testing time limits, a meta-analysis was determined to be the best source for an exploration of the effects of differences in time limitations (speededness) and test length (number of items) on underlying ability correlates of RAPM performance. Based on the considerations described up to this point, the main goal of the current investigation is to determine whether differences in the speededness and the length of the RAPM as implemented by various investigators results in differential construct validity (convergent and discriminant) for the RAPM. In particular, we hoped to address issues about whether speeded or short versions of the RAPM were more or less associated with general intelligence (IQ), fluid intelligence (*Gf*) at the broad trait level, were more or less associated with surface-level 'content' ability factors (i.e., from the spatial domain, given the figural content of the RAPM items), and were more or less associated with 'process' factors (namely factors associated with Perceptual Speed [PS]). Finally, even though we acknowledge there are far fewer studies in the literature that include correlations between RAPM and other abilities (e.g., verbal, numerical), where evaluation of discriminant validity can be directly assessed, we hoped to evaluate whether changes in speededness and length of the RAPM resulted in differences to correlations with traits that should indicate discriminant validity.

1.6. Abbreviated (short) test vs. standard test hypotheses

The general consensus regarding RAPM in the 'standard' version (where there are two worked examples in Set I, and a 5-min "practice" period completing the remaining 10 Set I items, followed by the 36-item Set II with a 40-min time limit) is that the Set II time limit is adequate to allow examinees to attempt as many items as they are likely to answer correctly, even if slowly and deliberately, except for guessing. In addition, there appear to be significant learning/'practice' effects both from the Set I experience and from the initial items on Set II, given the extant literature on practice effects associated with the RAPM (e.g., see Bors & Vigneau, 2003). Thus, ceteris paribus, presenting the examinee with abbreviated versions of the test will possibly impede the learning/practice effects.

There were many, sometimes competing, theoretical implications for differences in the underlying determinants of performance on abbreviated RAPM tests, in contrast to the 40-min Set II standard administration. On the one hand, the lack of practice could be construed to hypothesize that there would be an increase in the association between abbreviated RAPM scores and broad measures of intelligence, and similarly that there would be higher correlations between abbreviated RAPM scores and a wide variety of other ability measures, due to common method variance – that is, test sophistication or transfer (i.e., a decrease in the discriminant validity of the RAPM). On the other hand, if additional practice on the RAPM accorded under the 'standard' administration format reduces the impact of test familiarity and transfer, as examinees reach a point of common level of skill with the test format, then the 'standard' version should yield higher correlations with broad content (spatial) and general intellectual abilities. Because of these conflicting perspectives, we were agnostic about the direction of differences in correlations between abbreviated and standard administrations of the RAPM.

One key issue to be addressed regarding the abbreviated vs. standard versions of RAPM, however, was to evaluate how a failure to adjust the observed correlations for differences in test reliability, especially for the abbreviated tests, could result in underestimations of convergent validity and overestimates of discriminant validity, as a statistical artifact, based on the Spearman-Brown Prophecy Formula (e.g., see Brown, 1910; Spearman, 1904).

1.7. Speededness hypotheses

Our main conjecture regarding increasing the speededness of the RAPM (defined as an amount of time less than 36 items in 40 min or 66.66 s/item), is that performance on the RAPM would be more highly associated with measures of common 'content' (spatial abilities), because respondents will be forced to focus on more surface-level properties of the items. In addition, increasing speededness will also result in scores more highly associated with process/common-method measures (such as short-term memory, working memory, closure), and measures of perceptual speed that involve high memory demands (e.g., PS-Memory), compared to the standard administrations or even those with longer time-limit/item administrations than standard. It should be noted that predicted increases in correlations as a function of higher degree of speededness is independent of expected loss of reliability associated with shorter tests. Thus, we hypothesized the following:

Hypothesis 1. *Faster-than-standard administrations of the RAPM will be more highly correlated with Spatial abilities than standard or slower-than-standard administrations of the RAPM.*

Hypothesis 2. *Faster-than-standard administrations of the RAPM will be more highly correlated with PS-Memory abilities than standard or slower-than-standard administrations of the RAPM.*

Hypothesis 3. *Faster-than-standard administrations of the RAPM will be more highly correlated with Short-Term Memory and Working Memory abilities than standard or slower-than-standard administrations of the RAPM. (consistent with*Chuderski, 2013*).*

We also expected to find other patterns of results when comparing RAPM correlations under more or less speeded conditions, such as a lower discriminant validity with other abilities that are typically assessed using highly speeded tests. Because such results are potentially a result of common 'method' variance which may or may not be directly relevant to the underlying construct overlap, and because this issue was more exploratory, we did not develop a specific set of hypotheses for these results.

We cast a rather wide net to capture ability correlates of RAPM performance under varying speededness and test-length conditions. Although our main interest was in finding correlations between RAPM and various aspects of spatial, perceptual, and verbal abilities, we included all identifiable ability correlates in our initial review of the literature.

2. Method

2.1. Literature search

Our determination of study inclusion in the meta-analysis proceeded primarily in three stages: (1) Identification of short-form versions of RAPM; (2) Review of RAPM administrations; and (3) Final implementation of inclusion/exclusion criteria.

2.1.1. Identification of short-form versions of RAPM

We began our search by reviewing the literature to identify the most commonly used short-form versions of the RAPM. As briefly mentioned previously, RAPM is an unusual test because a number of different versions exist, but these various versions all use subsets of the original test items. The first version (Raven, Raven, & Court, 1962) of the full test consists of 48 items across two sets (12 items in Set I, 36 items in Set II) and roughly 50 min in total to administer, which is often inconvenient in a laboratory study context. This inconvenience has led to the development of several 'short-form' versions that typically consist of some subset of the original 36 Set II items administered under varying time constraints. Our search revealed four widely used short-form versions of RAPM, which are described below. It is important to note that this list is not exhaustive. In fact, other short forms of the test which consist of subsets of the original RAPM items were revealed in later stages of our search. However, we began with an exhaustive review of four widely used versions given their popularity in the field.

2.1.1.1. Arthur Jr. & Day (1994). The Arthur Jr. & Day RAPM short form consists of 12 items, selected to maintain the progressive nature that characterized the original RAPM (i.e., increasing item difficulty), and to maximize item-total correlations with Set II of the RAPM and the internal consistency reliability of the short form. Set II of the RAPM was divided into 12 sections of 3 items each that increased incrementally in difficulty (items 1–3, items 4–6, etc.). The item from each set that had the highest item-total correlation with Set II was selected for the short form. If two items had identical item-total correlations with Set II, the more difficult item was chosen. If two items had identical item-total correlations with Set II and identical difficulty, the item that enhanced the short form's internal consistency reliability was chosen. Although the article that accompanies the development of the Arthur and Day Jr. version made no claim nor recommendation regarding administering this short form under speeded conditions, other than offering the average time of completion, subsequent investigations by researchers who have used the Arthur Jr. and Day version have imposed a variety of time limits. The authors also did not provide directions for the number or content of items to administer to subjects as an introduction or practice prior to the scored items. Researchers who have subsequently administered this version often do not explicitly report the ways that participants are introduced to the test.

2.1.1.2. Bors and Stokes (1998). The Bors & Stokes RAPM short form consists of 12 items, selected to maximize the item-total correlations from Set II of the RAPM. When Bors and Stokes compared their version to the Arthurt Jr. & Day (1994) short form, the Bors and Stokes version had significantly lower performance, on the order of about 0.20 *sd* unit difference in means. (It should be noted, however, that the comparison was based on different samples of overlapping items from a single administration of the entire RAPM, so intercorrelations between 'independent' administrations of the two versions are not available.) These authors also noted that performance on their RAPM short form was not significantly affected by prior administration of the standard 12-item RAPM Set I when compared against the 'two instructional items' from Set I. The authors provided no indication of a time limit to be imposed in the short-form administration.

2.1.1.3. Hamel and Schmittmann (2006). The Hamel & Schmittmann "20-min" RAPM consists of the entire 36-item Set II of the standard RAPM, but administered under a 20-min time limit.

2.1.1.4. Odd-numbered items. A number of studies employed an 18-item version of RAPM in which only the odd-numbered items of the 36-item Set II are administered (e.g., see Burgoyne, Hambrick, & Altmann, 2019; Kane et al., 2004). With the other short-form versions, it was possible to find instances of their administration by reviewing the published articles which cited the article associated with the development of each. However, this 18-item version has no specific, published investigation associated with its development. Consequently, there was no simple way to provide an exact count of the frequency in which this version has been administered.

2.1.2. Review of RAPM administrations

We began by exhaustively scanning the abstracts of articles that cited each of the three named short-form versions, (745 abstracts according to the Google Scholar database as of December 10, 2020) and identified a list of articles that appeared to administer a short-form version of RAPM along with at least one test of an ability that we could identify from existing taxonomies (e.g., Carroll's, 1993, taxonomy), to an adult, non-clinical sample.

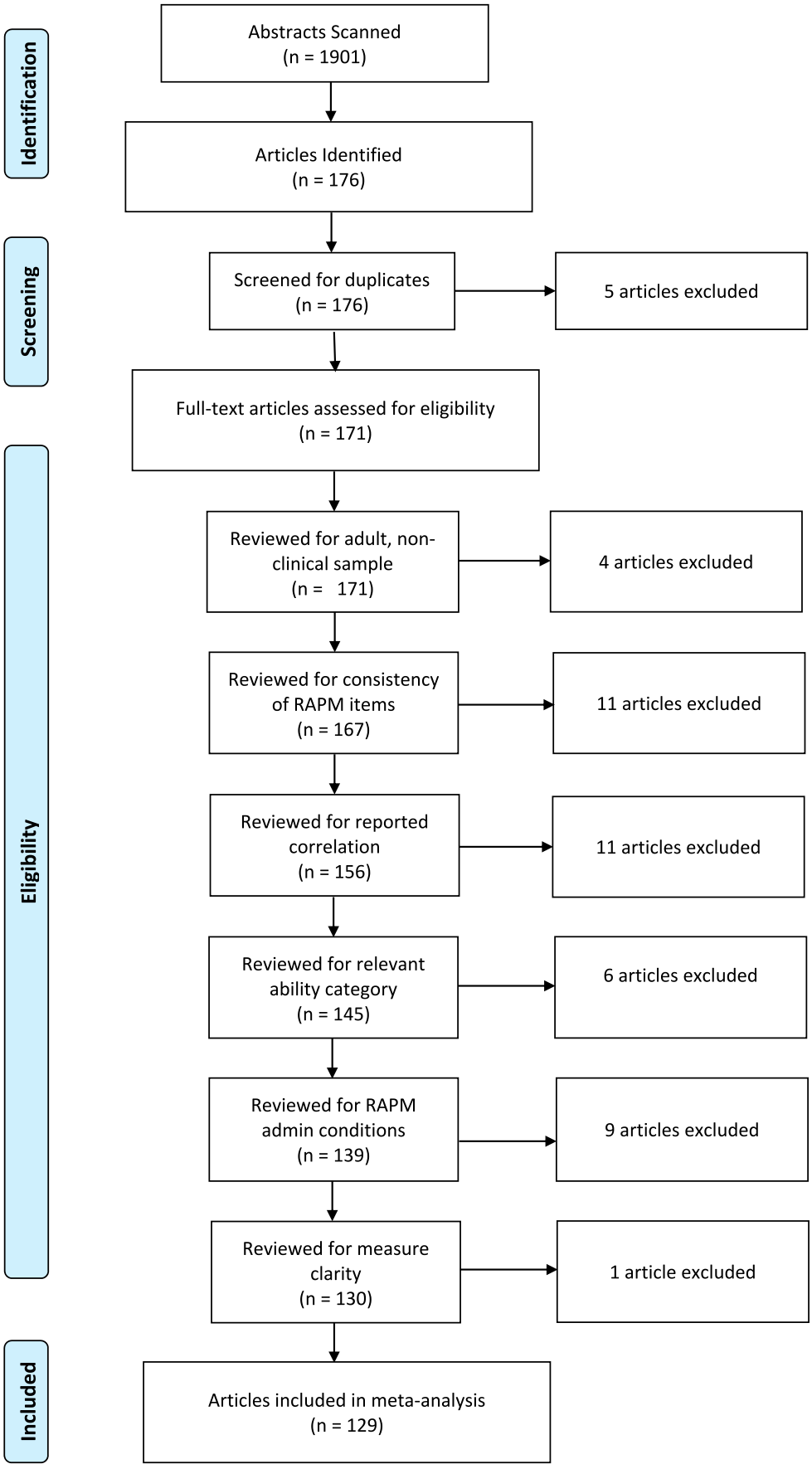
We then searched the Google Scholar database for additional studies that administered a version of RAPM along with tests of specific abilities included in Carroll's taxonomy using the following Boolean search terms: *"Raven's Advanced Progressive Matrices" AND "Spatial Ability", "Raven's Advanced Progressive Matrices" AND "Verbal Ability", "Raven's Advanced Progressive Matrices" AND "Numerical Ability", "Raven's Advanced Progressive Matrices" AND "Perceptual Speed", "Raven's Advanced Progressive Matrices" AND "Processing Speed"*. During this search, 1156 abstracts were scanned. Some studies identified in this stage included all 36 items from the standard version of the RAPM, while others used subsets of the original items.

2.1.2.1. Inclusion/Recalculation of studies from our laboratory. Finally, we included nine published studies and one unpublished study that were conducted in our lab. In the published studies, correlations between ability measures and the RAPM were computed and reported according to the participants' cumulative scores on RAPM Set I and Set II. However, the vast majority of the studies we reviewed computed and reported correlations only with Set II, even if they had administered Set I. Therefore, we re-computed the correlations from our studies, based only on the participants' performance on Set II of the RAPM.

2.1.3. Final implementation of inclusion/exclusion criteria

Of the 1901 abstracts produced by our search methodology, 176 included empirical investigations that administered some version of RAPM along with at least one additional ability test. These 176 articles were more thoroughly assessed based on a set of inclusion/exclusion criteria specified a priori. A visualization of this assessment process is provided Fig. 1. Five of the articles were not unique (e.g., dissertations that were later published as journal articles) and were excluded prior to assessment for eligibility. The remaining 171 papers were evaluated for eligibility based on the following criteria. The studies' samples were required be non-clinical and to consist of late adolescents or adults. Four articles that utilized child samples were excluded from analysis. Studies were required to contain a consistent set of RAPM items. In order to maintain consistency of the content of the test, only studies that used items from the 1962 version of the RAPM test (Raven et al., 1962) and more recent versions were included. Items from more recent versions were included because the items have not been altered in subsequent revisions. Eleven papers including studies that administered items that were either from versions prior to the 1962 version or from "non-advanced" versions of the Raven (e.g., Raven's Standard Progressive

Exclusion Flowchart



Note: This figure was constructed according to the PRISMA guidelines (see Moher et al., 2009).

Fig. 1. Inclusion/Exclusion Flowchart (Moher et al., 2009).

Matrices) were excluded from analysis.

Studies were also required to report at least one correlation between RAPM scores and another identifiable ability measure. Eleven papers that included studies that did not report correlations were excluded from analysis. At this point, we began a more thorough review of whether each of the remaining 145 articles administered a test that could be classified into a relevant ability category (within Carroll's, 1993 taxonomy). Each author separately reviewed a list of the individual tests that were administered in all 145 articles and sorted them into ability categories. Disagreements in categorization were resolved through discussion. 23 categories were identified: Spatial Ability, Spatial Visualization (Vz), Speeded Rotation (SR), Closure, Verbal Ability, Verbal/Fluency, Fluid Intelligence (Gf), Attention, Working Memory, Short-Term Memory, Memory, Learning, Creativity, Crystallized Intelligence (*Gc*), Knowledge, Mechanical Knowledge, Perceptual Speed, Perceptual Speed-Scanning (PS-Scanning), Perceptual Speed-Complex (PS-Complex), Perceptual Speed-Memory (PS-Memory), Perceptual Speed-Pattern Recognition (PS-Pattern Recognition), and Psychomotor. The **Appendix** presents representative tests from each category.[2] Six papers which only administered tests that could not be easily categorized under Carroll's taxonomy were excluded from analysis. Nine articles were removed because the conditions under which RAPM was administered (item length, time constraints, and/or item content) were unclear or otherwise unorthodox. Finally, one additional study was discarded because the nature of the ability test could not be determined. Based on these criteria, 55 articles were excluded – leaving a total of 129 articles comprised of 142 studies and 506 correlations computed over a total of 26,848 participants that were included in the meta-analysis.

2.1.3.1. Sources for Meta-Analysis. The following references were included in our meta-analysis: Ackerman (1986, 1988, 1990, 1992, 2000), Ackerman & Beier (2007), Ackerman et al. (2002), Ackerman & Kanfer (1993), Ackerman & Wolman (2007), Alberts (2007), Allan (2018), Arthur Jr. & Day (1991), Babcock (1994), Babcock & Laguna (1996), Baghaei, Khoshdel-Niyat, & Tabatabaee-Yazdi (2017), Batey, Furnham, & Safiullina (2010), Birney et al. (2017), Bruza, Welsh, & Navarro (2008), Buckley, Seery, Canty, & Gumaelius (2018), Burgoyne, Hamrick et al. (2019), Burgoyne, Harris, & Hambrick (2019), Chiesi, Ciancaleoni, Galli, Morsanyi, & Primi (2012), Choi & L'Hirondelle (2005), Chow (2017), Chuderski (2013), Cockcroft & Israel (2011), Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero (2012), Colom, Escorial, & Rebollo (2004), Colom, Rebollo, Palacios, Juan-Espinosa, & Kyllonen (2004), Coyle & Pillow (2008), Culbertson, Huffcutt, & Goebl (2013), Dang, Braeken, Ferrer, & Liu (2012), Darowski, Helder, Zacks, Hasher, & Hambrick (2008), De Simoni & von Basian (2018), DeYoung, Peterson, & Higgins (2005), Dodonova & Dodonov (2012), Edwards (2004), Embretson (1998), Ettinger & Corr (2001), Felez-Nobrega, Foster, Puig-Ribera, Draheim, & Hillman (2018), Furlan (2011), Furlan, Agnoli, & Reyna (2016), Graham (2011), Greengross & Miller (2011), Griffin, Carless, & Wilson (2013), Grounds (2016), Gutierrez et al. (2018), Guye & von Bastian (2017), Haavisto & Lehto (2005), Haier, Siegel, Tang, Abel, & Buchsbaum (1992), Hancock (2017), Hannon (2016), Hannon & Daneman (2014), Hardmeier & Schwaninger (2008), Hicks, Foster, & Engle (2016), Huettig & Janse (2016), Hunt, Pellegrino, Frick, Farr, & Alderton (1988), Israel (2006), Jaeggi et al. (2010), Jarosz & Wiley (2012), Kaesler, Welsh, & Semmler (2016), Kane et al. (2004), Kaufman, DeYoung, Gray, Brown, & Mackintosh (2009), Kaufman, DeYoung, Reis, & Gray (2011), Kock & Schlechter (2009), Koenig, Frey, & Detterman (2008), Kpolovie & Emekene (2016), Kranzler & Jensen (1991), Kulikowski & Orzechowski (2019), Lee & Therriault (2013), Li, Ren, Schweizer, Brinthaupt, & Wang (2021), Liberali, Reyna, Furlan, Stein, & Pardo (2012), Lilienthal, Tamez, Myerson, & Hale (2013), Lin, Hsu, Chen, & Wang (2012), Mackintosh & Bennett (2003, 2005), Martin, Mashburn, & Engle (2020), Martinez (2019), McCrory & Cooper (2007), McPherson & Burns (2007, 2008), McRorie & Cooper (2003,2004a, 2004b), Mellers et al. (2015), Miroshnik & Shcherbakova (2019), Morsanyi, Handley, & Serpell (2013), Morsanyi, O'Mahony, & McCormack (2017), Naber (2015), Neubauer & Bucik (1996), Oswald, McAbee Redick, & Hambrick (2015), Pahor, Stavropoulos, Jaeggi, & Seitz (2019), Park & Cho (2019), Paul (1986), du Pont et al. (2020), Prasad (2014), Ren, Schweizer, Wang, Chu, & Gong (2017), Ren, Tong, Peng, & Wang (2020), Richmond (2015), Rohde (2008), Rohde & Thompson (2007), Saccuzzo, Craig, Johnson, & Larson (1996), Sanchez et al. (2010), Schwarb (2012), Sefcek & Figueredo (2010), Seidler (2014), Sevenants, Dieussaert, & Schaeken (2013), Shelton, Elliot, Matthews, Hill, & Gouvier (2010), Siebert (2019), Singh, Gignac, Brydges, & Ecker (2018), Skagerlund, Forsblad, Slovic, & Västfjäll (2020), Srisang (2017), Stanovich & Cunningham (1992), Tabatabaee-Yazdi & Baghaei (2018), Tabe (2019), Teunisse, Case, Fitness, & Sweller (2020), Villado, Randall, & Zimmer (2016), Wang (2012), Waschl, Nettlebeck, & Burns (2017), Wei, Yuan, Chen, & Zhou (2012), Williams & Pearlberg (2006), Winman, Juslin, Lindskog, Nilsson, & Kerimi (2014), Xie (2015), Zajenkowski & Szymanik (2013), Zajenkowski, Stolarski, Maciantowicz, Malesza, & Witowska (2016), Zhu et al. (2010), Zimowski & Wothke (1988), and Zmigrod & Zmigrod (2016).

2.2. Operationalization of speed and length of RAPM administration

In order to investigate the influence of both time constraints and test length on the construct validity of the RAPM, it was necessary to operationalize various levels of both speededness and test length**.** In order to operationalize speededness, we sorted each study into three levels of time constraints based on the seconds allowed per item. The manual for the original RAPM specifies that the 36 items in Set II should be administered with a time limit of 40 min (i.e., 66.66 s/item) (Raven et al., 1962). Therefore, we classified any test that allowed 60–79 s/item as "Standard administration", any test that required fewer than 60 s/item as "Faster than standard administration", and any test allowed more than 79 s/item (including unspeeeded/power tests) as "Slower than standard administration". See Fig. 2 for a visualization of the time limits imposed in the studies included in our analyses.

In order to operationalize test length, we sorted each study into three categories: "Short RAPM version", "Medium RAPM version", and "Long RAPM version". The cutoff values for each category were determined by the most frequent item-lengths represented in our database. Fig. 3

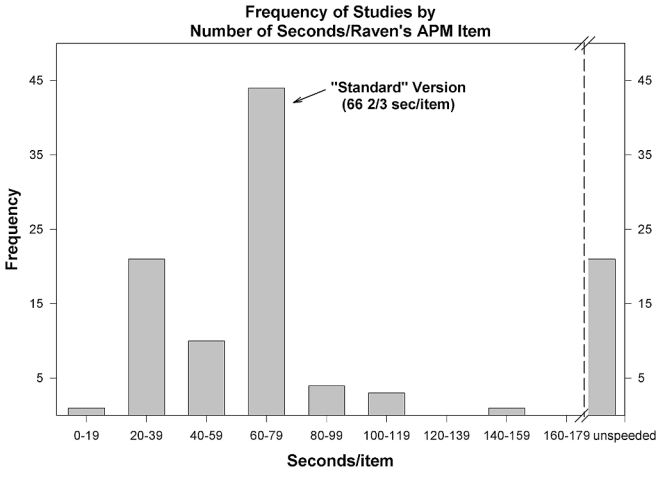


Fig. 2. Frequency of studies by number of seconds/Raven item.

[2] Prior to analyses, six categories were excluded because they contained a small number of studies that were largely administered under similar conditions of both timing and length. These six categories were Spatial Ability, Creativity, Perceptual Speed, PS-Complex, and Mechanical Knowledge. The categories are not included in the **Appendix.**

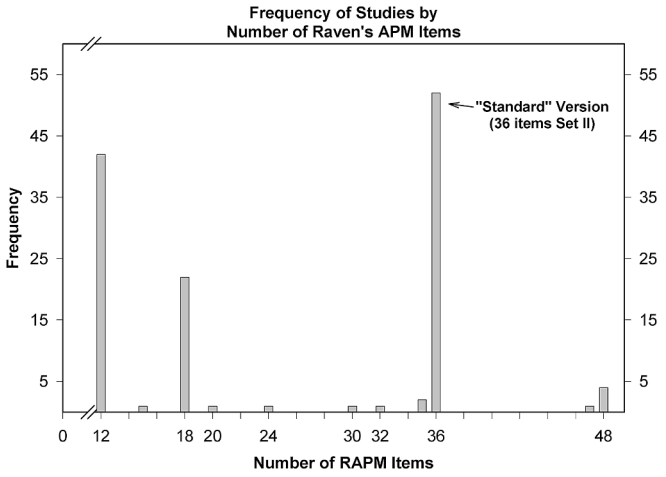


Fig. 3. Frequency of Studies by Number of Raven Items.

presents a plot of the frequency of item length for the 142 studies included in the meta-analysis. Based on the local maxima in Fig. 3, RAPM versions with 12 items were classified as "Short", RAPM versions with 13–35 items were classified as "Medium", and RAPM versions with at a least 36 items were classified as "Long".

A violin plot of the RAPM administration for each study based on both time constraints and test length is shown in Fig. 4. There are a few notable features from the analysis of time constraints and test length, as follows: First, the studies that contained RAPM items with 12 items (i.e., "short administrations") reported the time limits imposed on participants less frequently than studies that contained RAPM versions with longer item lengths ($\chi^2 = 69.41, p < .01$, adjusted residual = 6.40). Next, the vast majority of 18-item RAPM versions (e.g., odd-numbered items) were administered under faster-than-standard time constraints (as reported previously, $\chi^2 = 69.41$ (6), $p < .01$, adjusted residual = 5.80). Finally, it appears that studies where the standard RAPM was administered (i.e., "long administrations") were more likely to administer the test under either standard or slower-than-standard time constraints, compared to studies that administered subsets of the RAPM items (as reported previously $\chi^2 = 69.41$ (6), $p < .01$, adjusted residual = 3.60).

2.3. Analytic strategy

To consolidate the data reported in the included studies, we applied meta-analytic techniques to derive estimated correlations between scores on RAPM and scores on ability tests across levels of time constraints and test length under which the RAPM was administered.

Sampling error and measurement error have been known to confound meta-analytic findings (Hunter & Schmidt, 1990). In order to account for sampling error, we computed meta-analytic estimates under

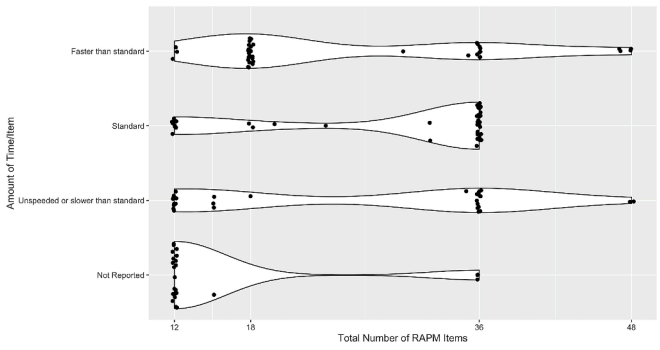


Fig. 4. Violin plot of speed and length of all administrations of RAPM included in meta-analysis.

the assumption of random-effects, which accounts for both within-study and between-study sampling error, rather than the assumptions of fixed-effects, which require an assumption that all studies consist of samples extracted from the same population (Borenstein, Hedges, & Rothstein, 2007). It has been established that random-effects models are less susceptible to Type-I error than fixed-effects models are (Hunter & Schmidt, 2000). The random-effects estimates were computed using the "metafor" *R* package (R Core Team, 2019; Viechtbauer, 2010). In this approach, Pearson's *r* coefficients are transformed via Fisher's *r*-to-*z* procedure to obtain a normal distribution. Next, the *z*-scores are weighted based upon sample size and variance, such that effect sizes from smaller studies with greater variance are weighted more heavily. Then the *z*-scores are aggregated and transformed back into *r* coefficients (Quintana, 2015). Within studies, correlations consisting of measures that assess the same ability (e.g., Paper Folding and Diagramming Relations both assess Vz) were aggregated to ensure that each sample is only represented a maximum of one time within each meta-analytic estimated effect size.

The second artifact that we partially accounted for was measurement error. In order to provide an estimate of the population correlation between the construct(s) underlying performance on the Raven's Progressive Matrices under different speed conditions and reference abilities, it is necessary to take account of the unreliability of the different measures. That is, ceteris paribus, a test with fewer items will provide a less reliable estimate of an underlying trait, when compared to the same test with more items.

The first issue to be addressed in estimating test reliability, however, is to decide *which* index of reliability is appropriate for the intended purpose. Thorndike's (1947) demarcation of the four sources of variance in test scores provides a useful framework for this determination. In Thorndike's framework, test score variance is categorized as two levels of permanence (temporary vs. lasting) and two levels of specificity (specific vs. general). For broad ability traits, the general consensus is that true-score variance will be in the quadrant identified as lasting and general, and that error-score variance is made up of the other three quadrants (temporary and general, temporary and specific, and lasting and specific). From this perspective, the most *appropriate* index of test reliability for such ability tests will be a delayed, alternate-form reliability index, as the delayed retest will average out the "temporary" sources of variance from the test-retest correlation and the alternate-form aspect will average out the 'specific" source of variance.[3] Unfortunately, we were unable to find any examples in the literature of any empirical study with delayed test-retest, alternate-form reliability indexes for the standard version of the Raven's Advanced Progressive Matrices test. This lack of empirical data is certainly attributable to the fact no alternate/parallel form of the Raven's Advanced Progressive Matrices test has been published.

In contrast, even though there are numerous estimates of internal consistency reliabilities (e.g., alpha, split-half, etc.) for the Raven, such estimates are not suitable for the current purposes (because they combine both temporary sources of item variance in the true-score variance category), and more generally, as noted by several investigators (e.g., Schmitt, 1996; Sijtsma, 2009), internal consistency estimates of reliability confound item homogeneity and reliability (e.g., narrow test content leads to higher internal consistency reliability estimates that may or may not have any bearing on the test-retest reliability estimates).

There are some estimates of delayed test-retest reliability for the 1940s version of the standard form of the Raven's Advanced Progressive

[3] It should be noted that even this approach does not entirely eliminate sources of variance in test scores that one *might* want to relegate to the error-score variance, such as the effects of familiarity with the test content and methods-of-working for test items – what one would typically refer to as 'learning' or 'practice' effects.

Matrices test. Early publications (where Set II had 48 items) had test-retest reliabilities reported on large samples by Vernon and Parry (1949) of $r_{xx'}$ = 0.79 ($N = 500$), 0.88 ($N = 500$), and 0.87 ($N = 1000$), but the tests were administered with a 20-min time limit (Vernon, 1947, Occup. Psych.). Raven, Raven, & Court (1998, Section 4) report a Set II (48-item) test-retest reliability of $r_{xx'} = 0.91$ from $N = 243$ "Adult Students," with the time limit of 40 min (p. APM7).

Only one study could be located with a delayed test-retest reliability with the 36-item RAPM Set II administered with the standard 40-min time limit (Bors & Vigneau, 2003). In a modest-sized sample ($N = 67$), diverse in age (from 26 to 79 years old), the authors reported a 45-day delayed test-retest reliability of $r_{xx'} = 0.85$.

For the 12-item APM short form with a 15-min administration time (Arthur & Day, 1994), the authors reported delayed test-retest reliability over one week of $r_{xx'} = 0.76$ ($N = 76$), and for 7–10 day interval, the test-retest reliability was $r_{xx'} = 0.76$ ($N = 111$) (see Arthur Jr., Tubre, Paul, & Sanchez-Ku, 1999).

In order to determine a rough estimate of reliability of the various item-length versions, even as we acknowledge the shortcoming of no adequate alternate forms of the RAPM, we used the above delayed test-retest correlations for the different item-length versions, and computed the estimated reliabilities from each correlation, using the Spearman-Brown Prophecy Formula as seen in (1) (Brown, 1910; Spearman, 1904).

$$\rho^{*}_{xx'} = \frac{n\rho_{xx'}}{1 + (n-1)\rho_{xx'}} \quad (1)$$

The appropriateness of this reliability estimation technique relies on the assumption that, due to error of measurement, correlations between observed scores (on psychological tests) will always be nearer to zero than would the corresponding correlations between so-called "true scores", in the classical test theory framework. Next, averages of actual and estimated test reliabilities were computed, with a mean correlation (after *r*-to-z transformation and subsequent *z*-to-*r* transformation). The estimated test-retest reliabilities were as follows: 12-item version ($r_{xx'}$ = 0.683), 18-item version ($r_{xx'}$ = 0.758), 20-item version ($r_{xx'}$ = 0.782), 24-tem version ($r_{xx'}$ = 0.811), 30-item version ($r_{xx'}$ = 0.843), 32-item version ($r_{xx'}$ = 0.851), 35-item version ($r_{xx'}$. = 862), 36-item version ($r_{xx'}$ = 0.865), 47-item version ($r_{xx'}$ 0.893), and 48-item version ($r_{xx'}$ = 0.895).

While it is generally accepted practice to correct for unreliability of both measures represented in a correlation, a majority of studies identified from our literature search did not report reliability estimates for the ability tests included in their analyses. Additionally, of the studies that did report reliability estimates, most reported internal consistency reliability, which would pose the same problems as described for RAPM. Accordingly, we chose not to adjust the reported correlations for unreliability of the ability measures correlated with RAPM scores. Therefore, our meta-analytic estimates reflect the estimated correlations between estimated true scores on the RAPM and observed scores on other ability tests.

3. Results

There are a number of salient findings from the meta-analysis of correlations between RAPM scores and scores on tests of other abilities, split out by speededness (in terms of *sec* allowed per item) and test length (in terms of the number of total items administered). **See**Tables 1 and 2 for the complete meta-analytic results.

3.1. Spatial visualization and RAPM

The first notable finding is related to the correlations of RAPM scores with tests of Spatial Visualization (Vz). Across levels of speededness and test length, five of the six estimated correlations between true scores on RAPM and observed scores on tests of Vz are greater than $\widehat{\rho} = 0.50$, meaning that VZ accounts for at least 25% of the variance in RAPM matrices for five of the six speed/length conditions. Given that the cutoff for a large correlation according to Cohen's (1988) criteria is 0.50,[4] this finding suggests that claims that RAPM measures "*g* and little else" (Jensen, 1980) are not justified and that performance on RAPM may be a function of other lower-order abilities, particularly Vz. However, the previous statement should be qualified by pointing out that the estimated relationships derived in this meta-analysis were corrected for unreliability in the RAPM but not in the other lower-order ability measures. This was due to the infrequency with which many of the articles reviewed in this effort reported reliability indices other than those representing internal consistency, which would not have been appropriate to use in corrections for unreliability.

Second, the estimated correlations between performance on RAPM and performance on tests of Vz rise noticeably under higher levels of time pressure ($\widehat{\rho}_{\text{slower, VZ}} = 0.436$, $\widehat{\rho}_{\text{faster, VZ}} = 0.698$), providing general support for Hypothesis #1. This suggests that changing the time constraints under which the RAPM is administered may alter the extent to which variance in performance is attributed to Vz. More specifically, scores on highly speeded RAPM versions may reflect an individual's Vz ability to a greater extent than scores on the RAPM administered under standard time constraints, slower than standard time constraints, or as an untimed power test. This suggests that researchers who justify the use of a speeded version of RAPM based solely on correlations between speeded and unspeeded versions of the RAPM may be underestimating the influence that particular abilities (e.g., Vz) have on performance under varying levels of time constraints. In short, the differences between these correlations indicate that time pressure may in fact result in a test of general reasoning or intellectual ability that draws on a different set of underlying cognitive processes, when compared to the same test administered with less stringent time constraints.

3.2. Perceptual speed - memory & RAPM

Another salient finding is that higher levels of speededness for the RAPM are associated with higher correlations between RAPM scores and Perceptual Speed (PS) ability measures, but most notably not with all PS abilities. That is, for PS-Memory tests (e.g., Digit/Symbol, Coding, Factors of 7 – see Ackerman & Cianciolo, 2000), scores on the faster-than-standard administration of RAPM were estimated to correlate $\widehat{\rho} = 0.515$ with PS-Memory, indicating a shared variance of roughly 25%, while standard administrations of RAPM correlated $\widehat{\rho} = 0.307$ (about 9% of shared variance), and slower than standard administrations correlated $\widehat{\rho} = 0.171$ (about 3% of shared variance), providing general support for Hypothesis #2. This pattern of results is also consistent with the notion that increasing the speededness of the RAPM increases demands on Short-Term Memory (STM)/Working Memory abilities – which was stated as Hypothesis #3 (and is consistent with Chuderski, 2013), and is also a common element for the PS-Memory tests. However, the lack of substantial correlations with the other PS abilities, in contrast to the PS-Memory construct, provides important qualifications to those investigations that have been perhaps cavalier about selecting reference measures for "Processing Speed" or "Perceptual Speed". A selection of PS-Scanning or PS-Pattern Recognition tests can be expected to result in lower communalities with RAPM scores (as illustrated in Tables 1 and 2), and thus a greater *relative* identification of RAPM with other abilities. In contrast, a selection of PS reference tests that involve PS-Memory, especially in the context of speeded RAPM administration will likely

[4] We thank an anonymous reviewer for pointing out that while Cohen's (1988) original criteria for defining a large, medium, and small correlation was somewhat subjective, other suggests that more lenient cutoffs should be used in individual differences research (e.g., Abelson, 1985; Gignac & Svadorai, 2016). While we rely on Cohen's traditional criteria, we offer this caveat and provide an estimate of the variance accounted for in hopes that readers may make their own decisions regarding the classification of correlation strength.

Table 1
Meta-analysis by speed.

	Raw			Corrected*		
	Slower or Unspeeded	Standard	Faster	Slower or Unspeeded	Standard	Faster
Spatial Vz						
$\widehat{\rho}$	*0.402*	**0.524**	**0.528**	*0.436*	0.**567**	**0.698**
CI	[0.368, 0.435]	[0.480, 0.565]	[0.397, 0.638]	[0.403, 0.468]	[0.521, 0.609]	[0.314, 0.886]
k	7	20	11	7	20	11
Q	3.60	38.99*	115.80*	4.97	46.85*	1404.73*
I^2	0%	52.33%	90.04%	0.14%	61.10%	99.06%
SR						
$\widehat{\rho}$	0.226	*0.319*	0.*420*	0.273	*0.345*	0.*483*
CI	[0.042, 0.395]	[0.253, 0.382]	[0.345, 0.489]	[0.093, 0.437]	[0.280, 0.406]	[0.483, 0.545]
k	1	7	4	1	7	4
Q	0.00	1.96	3.55	0.00	2.36	3.58
I^2	0%	0%	4.25%	0%	0%	0.02%
Closure						
$\widehat{\rho}$	*0.358*	*0.327*	0.**544**	0.*402*	0.*355*	0.**576**
CI	[0.184, 0.511]	[0.093, 0.528]	[0.453, 0.624]	[0.209, 0.565]	[0.124, 0.550]	[0.489, 0.651]
k	4	1	2	4	1	2
Q	11.36*	0.00	0.73	13.31*	0.00	0.91
I^2	83.32%	0%	0%	87.04%	0%	0%
Verbal						
$\widehat{\rho}$	0.*367*	0.297	0.*326*	0.*414*	0.*331*	0.*370*
CI	[0.283, 0.445]	[0.240, 0.351]	[0.219, 0.426]	[0.318, 0.501]	[0.265, 0.393]	[0.248, 0.480]
k	17	21	12	17	21	12
Q	73.24*	53.16*	81.35*	97.79*	77.49*	114.92*
I^2	85.88%	60.20%	87.08%	90.09%	71.47%	90.55%
Verbal/Fluency						
$\widehat{\rho}$	0.149	0.275	0.*316*	0.180	0.*300*	0.*356*
CI	[0.029, 0.265]	[0.187, 0.358]	[0.192, 0.429]	[0.061, 0.294]	[0.200, 0.394]	[0.206, 0.489]
k	1	10	5	1	10	5
Q	0.00	23.11*	13.94*	0.00	28.12*	21.42*
I^2	0%	69.93%	70.85%	0%	72.95%	80.89%
Math						
$\widehat{\rho}$	0.*420*	0.*382*	0.*420*	0.*478*	0.*425*	0.*458*
CI	[0.367, 0.470]	[0.311, 0.448]	[0.329, 0.503]	[0.413, 0.538]	[0.342, 0.502]	[0.347, 0.557]
k	11	13	4	11	13	4
Q	17.45	37.51*	5.56	26.75*	57.90*	7.80
I^2	39.27%	68.00%	42.39%	63.83%	78.54%	62.68%
IQ						
$\widehat{\rho}$	0.*490*	0.*409*	0.**591** [	0.**554**	0.*485*	0.**685**
CI	[0.398, 0.572]	[0.220, 0.568]	0.533, 0.644]	[0.443, 0.649]	[0.233, 0.675]	[0.608, 0.749]
k	9	3	3	9	3	3
Q	35.15*	19.06*	2.88	51.98*	36.95*	5.86
I^2	78.15%	86.52%	23.32%	87.16%	93.10%	68.05%
Gf						
$\widehat{\rho}$	0.*481*	0.*485*	0.*442*	0.**548**	0.**534**	0.**502**
CI	[0.368, 0.580]	[0.428, 0.537]	[0.395, 0.486]	[0.425, 0.651]	[0.474, 0.589]	[0.450, 0.550]
k	13	28	22	13	28	22
Q	159.93*	250.86*	75.46*	215.44*	307.10*	106.70*
I^2	93.17%	82.73%	72.12%	94.87%	85.93%	80.17%
Attention						
$\widehat{\rho}$		0.233	0.265		0.257	0.296
CI		[−0.095, 0.516]	[0.186, 0.342]		[−0.090, 0.548]	[0.206, 0.381]
k	No studies	2	7	No Studies	2	7
Q		7.28*	11.65		8.16*	15.07*
I^2		86.25%	49.14%		87.75%	61.02%
WM						
$\widehat{\rho}$	0.258	0.*300*	0.*319*	0.*305*	0.*336*	0.*360*
CI	[0.197, 0.317]	[0.248, 0.350]	[0.275, 0.362]	[0.235, 0.371]	[0.280, 0.390]	[0.309, 0.410]
k	8	13	23	8	13	23
Q	7.44	15.60	50.47*	10.75	19.27	76.21*
I^2	20.88%	27.08%	56.96%	39.71%	39.56%	70.60%
STM						
$\widehat{\rho}$	0.241	0.*478*	0.*336*	0.*315*	0.**578**	0.*373*
CI	[0.054, 0.413]	[0.273, 0.641]	[0.290, 0.380]	[0.066, 0.527]	[0.397, 0.716]	[0.325, 0.419]
k	3	1	7	3	1	7

(*continued on next page*)

Table 1 (*continued*)

	Raw			Corrected*		
	Slower or Unspeeded	Standard	Faster	Slower or Unspeeded	Standard	Faster
Q	3.92	0.00	6.09	6.46*	0.00	6.71
I^2	48.59%	0%	0%	70.67%	0%	12.23%
Memory						
$\widehat{\rho}$	0.297	0.*323*	0.*338*	0.*327*	0.*357*	0.*411*
CI	[0.191, 0.362]	[0.255, 0.387]	[0.316, 0.457]	[0.187, 0.455]	[0.280, 0.429]	[0.339, 0.477]
k	4	9	1	4	9	1
Q	6.91	11.31	0.00	11.58*	14.29	0.00
I^2	58.74%	24.50%	0%	84.37%	42.88%	0%
Learning						
$\widehat{\rho}$	0.070	0.250	0.271	0.075	0.272	0.310
CI	[−0.081, 0.217]	[0.155, 0.340]	[0.096, 0.430]	[−0.075, 0.222]	[0.166, 0.371]	[0.103, 0.491]
k	1	5	5	1	5	5
Q	0.00	4.49	34.04*	0.00	5.48	52.30*
I^2	0%	4.52%	83.75%	0%	23.58%	88.83%
Gc						
$\widehat{\rho}$	0.*336*	0.294	0.*438*	0.*362*	0.*318*	0.**503**
CI	[0.151, 0.499]	[0.225, 0.360]	[0.327, 0.537]	[0.179, 0.520]	[0.245, 0.387]	[0.400, 0.594]
k	1	5	1	1	5	1
Q	0.00	4.50	0.00	0.00	5.43	0.00
I^2	0%	0%	0%	0%	10.11%	0%
Knowledge						
$\widehat{\rho}$	0.264	0.*417*	0.124	0.279	0.**520**	0.137
CI	[0.135, 0.383]	[−0.017, 0.718]	[−0.017, 0.260]	[0.151, 0.397]	[−0.054, 0.836]	[−0.017, 0.284]
k	1	2	2	1	2	2
Q	0.00	30.89*	1.54	0.00	57.81*	1.80
I^2	0%	96.76%	35.27%	0%	98.27%	44.41%
PS-Scanning						
$\widehat{\rho}$	0.114	0.226	0.268	0.142	0.291	0.295
CI	[−0.005, 0.229]	[0.207, 0.324]	[0.165, 0.366]	[0.002, 0.277]	[0.229, 0.352]	[0.184, 0.400]
k	4	19	13	4	19	13
Q	3.40	39.92*	74.25*	4.99	46.02*	88.69*
I^2	22.74%	54.17%	83.65%	41.69%	59.93%	86.27%
PS-Memory						
$\widehat{\rho}$	0.157	0.270	0.*458*	0.171	0.*307*	0.**515**
CI	[−0.017, 0.321]	[0.192, 0.344]	[0.390, 0.521]	[−0.017, 0.347]	[0.219, 0.391]	[0.438, 0.585]
k	2	8	4	2	8	4
Q	2.01	9.97	4.60	2.35	13.44	5.19
I^2	50.24%	31.83%	0.01%	57.39%	48.77%	27.27%
PS-Pattern						
$\widehat{\rho}$	0.245	0.226	0.*314*	0.263	247	0.*349*
CI	[0.011, 0.454]	[0.126, 0.321]	[0.124, 0.482]	[0.030, 0.469]	[0.139, 0.348]	[0.125, 0.539]
k	1	11	3	1	11	3
Q	0.00	31.52*	7.88*	0.00	36.77*	10.69*
I^2	0%	68.39%	81.49%	0%	73.17%	87.01%
Psychomotor						
$\widehat{\rho}$	0.178	0.214	0.207	0.191	0.231	0.223
CI	[−0.059, 0.396]	[0.140, 0.286]	[−0.013, 0.408]	[−0.046, 0.408]	[0.157, 0.302]	[0.003, 0.422]
k	1	7	1	1	7	1
Q	0.00	3.35	0.00	0.00	3.93	0.00
I^2	0%	0%	0%	0%	0%	0%

Estimations in cells with k = 1 were calculated using fixed effects rather than random effects. 95% confidence intervals are indicated in brackets. Estimated correlations with a value, at or greater than, 0.500 are **bolded**, those with a value, at or greater than 0.300 and less than 0.500 are *italicized*, and those with a value less than 0.300 are in plain text. k = number of correlations in the meta-analysis. $\widehat{\rho}$ = estimated correlation in the population. Q = statistical test of the heterogeneity of component correlations in each cell (those with an * are significant at the $p < .05$ level. I^2 = estimation of the proportion of variance that is due to heterogeneity among the component correlations in each cell. Spatial Vz = Spatial Visualization, SR = Speeded Rotation, Gf = Fluid Intelligence, WM = Working Memory, STM = Short-Term Memory, Gc = Crystallized Intelligence. PS = Perceptual speed. Corrected* = Corrected for RAPM unreliability only.

result in findings of greater communality with 'PS ability' – see for example, the raw and partial correlations between PS abilities and *g* in Ackerman et al. (2002).

Additionally, a comparison of correlations between RAPM scores and tests of abilities that are corrected for unreliability and not corrected for unreliability reveal an interesting trend. The most noticeable differences between correlations that are corrected vs. not corrected occur under faster-than-standard speed administrations and when RAPM measures contain fewer items. The projected reduction of reliability associated with tests of fewer items was entirely expected, given it was derived from the Spearman-Brown formula. Generally, tests with shorter administration times tend to be less reliable than longer administration times, but there is no deterministic formula for projecting a particular loss function for such comparisons. If the correlations reported in the original articles were to be corrected based on estimates of test-retest reliability, the estimated correlations would likely be even higher than

Table 2
Meta-analysis by item length.

	Raw			Corrected*		
	Short	Medium	Long	Short	Medium	Long
Spatial Vz						
$\widehat{\rho}$	0.*411*	0.**536**	0.**515**	0.**500**	0.**702**	0.**559**
CI	[0.311, 0.502]	[0.416, 0.639]	[0.477, 0.551]	[0.373, 0.608]	[0.358, 0.878]	[0.519, 0.597]
k	7	12	24	7	12	23
Q	12.84*	123.74*	65.82*	24.23*	1420.89*	83.40*
I^2	55.10%	89.78%	57.89%	75.97%	99.00%	67.59%
SR						
$\widehat{\rho}$	0.*329*	0.275	0.*343*	0.*391*	0.*308*	0.*375*
CI	[0.198, 0.447]	[0.116, 0.421]	[0.247, 0.432]	[0.236, 0.527]	[0.151, 0.450]	[0.284, 0.460]
k	3	2	7	3	2	8
Q	3.60	0.48	13.68*	5.93	0.74	17.11*
I^2	46.27%	0%	59.51%	63.66%	0%	60.65%
Closure						
$\widehat{\rho}$	0.262	0.279	0.**512**	0.*317*	0.*301*	0.**548**
CI	[0.190, 0.331]	[0.165, 0.386]	[0.369, 0.631]	[0.247, 0.384]	[0.188, 0.407]	[0.395, 0.672]
k	2	2	4	2	2	4
Q	0.30	0.24	9.22*	0.47	0.30	11.27*
I^2	0%	0%	69.19%	0%	0%	75.31%
Verbal						
$\widehat{\rho}$	0.271	0.*377*	0.*319*	0.*338*	0.*437*	0.*346*
CI	[0.205, 0.334]	[0.283, 0.463]	[0.261, 0.374]	[0.251, 0.420]	[0.321, 0.540]	[0.282, 0.406]
k	14	13	33	14	13	33
Q	38.14*	72.89*	127.16*	63.63*	117.67*	155.49*
I^2	67.27%	84.84%	81.81%	82.82%	91.05%	85.65%
Verbal/Fluency						
$\widehat{\rho}$	0.149	0.*344*	0.266	0.180	0.*391*	0.290
CI	[0.029, 0.265]	[0.216, 0.460]	[0.184, 0.343]	[0.061, 0.294]	[0.237, 0.527]	[0.197, 0.377]
k	1	4	11	1	4	11
Q	0.00	10.61*	24.00*	0.00	16.35*	29.19*
I^2	0%	71.56%	59.95%	0%	81.48%	70.02%
Math						
$\widehat{\rho}$	0.*345*	0.*410*	0.*385*	0.*418*	0.*470*	0.*414*
CI	[0.282, 0.406]	[0.355, 0.462]	[0.327, 0.440]	[0.338, 0.493]	[0.415, 0.522]	[0.351, 0.472]
k	19	5	18	19	5	18
Q	55.11*	2.45	48.57*	94.10*	4.66	58.44*
I^2	71.93%	0%	66.20%	84.75	13.75%	72.31%
IQ						
$\widehat{\rho}$	0.*428*	0.*495*	0.**530**	0.**528**	0.**572**	0.**564**
CI	[0.345, 0.504]	[0.343, 0.621]	[0.455, 0.598]	[0.415, 0.626]	[0.385, 0.714]	[0.477, 0.640]
k	9	5	6	9	5	6
Q	34.82*	35.36*	8.78	69.39*	61.36*	11.26*
I^2	77.69%	89.56%	39.59%	90.49%	94.07%	57.99%
Gf						
$\widehat{\rho}$	0.*404*	0.*421*	0.**518**	0.*495*	0.*481*	0.**560**
CI	[0.352, 0.454]	[0.374, 0.466]	[0.464, 0.569]	[0.431, 0.553]	[0.425, 0.533]	[0.501, 0.615]
k	17	20	32	17	20	32
Q	35.56*	51.79*	442.08*	62.52*	83.69*	596.35*
I^2	55.67%	63.17%	89.46%	75.06%	76.85%	92.32%
Attention						
$\widehat{\rho}$	0.070	0.261	0.*320*	0.085	0.296	0.*341*
CI	[−0.096, 0.232]	[0.187, 0.332]	[0.181, 0.446]	[−0.081, 0.246]	[0.210, 0.378]	[0.187, 0.479]
k	1	6	3	1	6	3
Q	0.00	7.72*	4.13	0.00	10.78	4.80
I^2	0%	35.41%	51.81%	0%	53.62%	61.50%
WM						
$\widehat{\rho}$	0.261	0.*326*	0.290	0.*316*	0.*373*	0.*315*
CI	[0.213, 0.309]	[0.272, 0.379]	[0.249, 0.330]	[0.258, 0.371]	[0.310, 0.433]	[0.271, 0.357]
k	12	17	19	12	17	19
Q	15.06	41.72*	23.36	23.73*	61.58*	28.02
I^2	32.11%	61.08%	28.06%	54.36%	72.99%	38.41%
STM						
$\widehat{\rho}$	0.*303*	0.*309*	0.*335*	0.*375*	0.*356*	0.*360*
CI	[0.160, 0.434]	[0.236, 0.379]	[0.257, 0.409]	[0.200, 0.527]	[0.273, 0.433]	[0.278, 0.436]
k	6	4	4	6	4	4

(*continued on next page*)

Table 2 (*continued*)

	Raw			Corrected*		
	Short	Medium	Long	Short	Medium	Long
Q	26.72*	3.52	5.14	43.34*	5.04	5.77
I^2	78.70%	22.62%	42.09%	87.13%	42.57%	48.37%
Memory						
$\widehat{\rho}$	0.*371*	0.*361*	0.*310*	0.*456*	0.*398*	0.*334*
CI	[0.197, 0.523]	[0.244, 0.469]	[0.249, 0.368]	[0.244, 0.626]	[0.283, 0.501]	[0.265, 0.399]
k	3	3	9	3	3	9
Q	9.91*	0.03	18.11*	16.58*	0.07	21.19*
I^2	77.91%	0%	55.49%	86.42%	0%	66.13%
Learning						
$\widehat{\rho}$	0.178	0.*379*	0.148	0.215	0.*425*	0.158
CI	[−0.097, 0.428]	[0.217, 0.521]	[0.089, 0.206]	[−0.058, 0.459]	[0.234, 0.585]	[0.099, 0.216]
k	1	4	6	1	4	6
Q	0.00	11.34*	3.62	0.00	17.79*	4.27
I^2	0%	70.12%	0%	0%	79.59%	0%
Gc						
$\widehat{\rho}$	0.291	0.*438*	0.299	0.*352*	0.**503**	0.*322*
CI	[0.143, 0.427]	[0.327, 0.537]	[0.235, 0.361]	[0.209, 0.481]	[0.400, 0.594]	[0.258, 0.383]
k	1	1	6	1	1	6
Q	0.00	0.00	4.69	0.00	0.00	5.66
I^2	0%	0%	0.04%	0%	0%	0.09%
Knowledge						
$\widehat{\rho}$	0.*417*		0.165	0.**520**		0.177
CI	[−0.017, 0.718]		[0.077, 0.250]	[−0.054, 0.836]		[0.085, 0.266]
k	2	no studies	4	2	no studies	4
Q	38.89*		4.98	57.81*		5.56
I^2	96.76%		41.67%	98.27%		47.60%
PS-Scanning						
$\widehat{\rho}$	0.137	0.181	0.*317*	0.177	0.209	0.*340*
CI	[0.050, 0.221]	[0.097, 0.263]	[0.252, 0.375]	[0.067, 0.282]	[0.112, 0.302]	[0.273, 0.404]
k	7	8	22	7	8	22
Q	9.07	15.38*	69.09*	13.76*	20.85*	82.92*
I^2	35.16%	55.13%	67.73%	59.05%	66.54%	72.83%
PS-Memory						
$\widehat{\rho}$	0.212	0.*434*	0.280	0.266	0.*490*	0.*304*
CI	[0.019, 0.389]	[0.357, 0.505]	[0.160, 0.391]	[0.036, 0.469]	[0.397, 0.573]	[0.173, 0.424}
k	4	3	8	4	3	8
Q	11.20*	2.31	28.14*	17.03*	3.43	34.10*
I^2	70.18%	0.11%	76.28%	79.56%	33.53%	80.59%
PS-Pattern						
$\widehat{\rho}$	0.031	0.237	0.297	0.038	0.268	0.*320*
CI	[−0.071, 0.132]	[−0.004. 0.451]	[0.220, 0.370]	[−0.064, 0.139]	[−0.013, 0.510]	[0.237, 0.399]
k	2	3	10	2	3	10
Q	0.22	9.58*	17.36*	0.32	13.64*	20.66*
I^2	0%	76.37%	48.82%	0%	82.89%	57.28%
Psychomotor						
$\widehat{\rho}$	0.110	0.245	0.207	0.133	0.265	0.223
CI	[−0.261, 0.452]	[0.012, 0.452]	[0.137, 0.276]	[−0.239, 0.470]	[0.034, 0.470]	[0.153, 0.291]
k	1	1	8	1	1	8
Q	0.00	0.00	3.34	0.00	0.00	3.91
I^2	0%	0%	0%	0%	0%	0%

Estimations in cells with k = 1 were calculated using fixed effects rather than random effects. 95% confidence intervals are indicated in brackets. Estimated correlations with a value, at or greater than, 0.500 are **bolded**, those with a value, at or greater than 0.300 and less than 0.500 are *italicized*, and those with a value less than 0.300 are in plain text. k = number of correlations in the meta-analysis. $\widehat{\rho}$ = estimated correlation in the population. Q = statistical test of the heterogeneity of component correlations in each cell (those with an * are significant at the $p < .05$ level. I^2 = estimation of the proportion of variance that is due to heterogeneity among the component correlations in each cell. Spatial Vz = Spatial Visualization, SR = Speeded Rotation, Gf = Fluid Intelligence, WM = Working Memory, STM = Short-Term Memory, Gc = Crystallized Intelligence. PS = Perceptual speed. Corrected* = Corrected for RAPM unreliability only.

the raw average correlations indicate. The vast majority of the studies we reviewed did not report any corrections for reliability. Given this information, it is possible that investigations using a short, speeded version of RAPM that does not correct for unreliability contains higher estimated true score correlations between RAPM scores and scores on tests of Vz or PS-Memory than results suggest and also higher correlations with other abilities (i.e., less discriminant validity for the RAPM).

When estimated correlations were separated based on RAPM test length, many of the correlations with particular abilities displayed a curvilinear trend in which correlations under medium length RAPM measures (i.e., 13–35 items) were higher than correlations under short (i.e., 12 or fewer items) or long (i.e., 36 or more items). This trend can be observed for correlations between RAPM and tests that measure Vz, Verbal, Math, and Working Memory abilities. While this finding was

Table 3
Crosstabs of frequency of studies by number of seconds/Raven item speed and length of all administrations of RAPM included in meta-analysis.

	Not reported	Unspeeded or slower	Standard	Faster than Standard	Total
Short	22	10	11	4	47
Medium	0	5	6	22	33
Long	3	15	31	13	62
Total	25	30	48	39	142

initially unexpected, it can be reconciled by integrating the trends based on speed with the trends displayed by Table 3 and Fig. 4. As illustrated in Table 3 and Fig. 4, RAPM versions consisting of 18 items (which are all categorized as "medium" length) were more consistently administered under faster-than-standard time conditions (as reported previously, $\chi^2 = 69.41$ (6), $p < .01$, adjusted residual = 5.80).

The curvilinear trends in correlations between RAPM performance and ability test performance in these ability categories may reflect the fact that 18-item versions were often administered under greater time constraints, which further indicates a change in cognitive processing under time pressure.

While investigating the relationship between RAPM performance and Working Memory (WM) ability was not a direct aim of the present work, many of the studies included in our literature search presented correlations between measures of WM and RAPM. Within each level of speededness and test length, performance on RAPM was moderately correlated with measures of WM (estimated correlations ranging from $\widehat{\rho} = 0.32–0.37$). This range is well beneath those observed in the early work which spurred the claim that WM and *Gf* are nearly indistinguishable constructs (Kyllonen & Christal, 1990). The range is, however, in concordance with the low end of a previous a meta-analytic effort which presented estimated correlations between WM and Gf measures ranging from $\bar{r} = 0.30$ to 0.80 (Chuderski, 2013). Additionally, the finding that the correlations between WM and RAPM presented here increase as the speededness of RAPM is increased is generally consistent with the previous observation that WM becomes more highly related to *Gf* as time pressure is applied during administration of measures of *Gf* (Chuderski, 2013). We do not agree with the 'isomorphic' conclusion under any of the administrative conditions we examined, given the modest magnitudes of these correlations. It is worth noting that claims regarding the similarities between gf and WM are often made at the level of latent constructs, whereas the relationships reviewed in this paper are at the level of specific tests and/or specific tasks.

Finally, it is also noteworthy that the relationships between RAPM performance and WM were similar in magnitude to those between RAPM and STM measures across all RAPM administration formats. One interpretation of these findings is that the processing/attentional component (which has been hypothesized to distinguish STM from WM – see for example Engle, 2002) plays a minimal role in performance on RAPM, which reinforces the contention that WM is a *different* construct from *Gf* (Ackerman, Beier, & Boyle, 2005).

4. Discussion

Overall, the meta-analytic results appeared to support our conjectures that increasing RAPM speededness and/or reducing the number of items administered in the test results in differences in the common variance among various abilities and RAPM performance. These findings point to the potential for greater or lesser influences of "content" abilities (e.g., Vz) and process-oriented abilities (PS-Memory, STM, WM), depending on what appeared to result from rather mundane desires to make RAPM testing more efficient in laboratory studies. Because of the somewhat sparse matrix of studies for several ability/RAPM correlations, that is, multiple cells with few independent studies and thus wide confidence intervals, these results are suggestive, but not definitive, for many ability/RAPM combinations.

In particular, scores on faster-than-standard administrations of the RAPM appear to more highly saturated with Vz, PS-Memory, and STM/WM abilities, making the RAPM less likely to be identifiable as a 'pure' measure of *g* or *Gf*. Whether using a faster-than-standard RAPM test as a sole indicator of g/*Gf* results in 'process contamination' for other measures correlated with the RAPM may well depend on the speededness demands of the other measures (e.g., see Fry & Hale, 1996), but the additional influence of such method factors cannot be discounted, based on the current meta-analytic results.

A more fundamental question relates to the relationship between RAPM and intelligence at a more general level. In the original instantiation of the Progressive Matrices Test (Penrose & Raven, 1936) and subsequent standard versions of the test, there is an implied high degree of overlap between the Raven test scores and the construct of general intelligence. But, the administration instructions appear to insure that examinees are not pushed to rapidly complete the test, rather their performance is mainly limited by the power-test format, and not directly be the speed of responses. Thus, one should be able to reasonably infer that speed of responding is not a major characteristic of the ability construct underlying performance on the RAPM. However, this perspective is at odds with some theoretical views of intelligence (e.g., Thorndike et al., 1926) and the vast majority of research that has been inspired by the information processing approach to intelligence since the 1970s (e.g., Hunt, Frost, & Lunneborg, 1973). So, depending on a researcher's definition of intelligence as more or less related to determining the correct answer "quickly", versions of the RAPM that introduce a speededness component may in fact be viewed as more representative of the construct of intelligence than a version administered with few or minimal time limits. Based on the current meta-analysis results, though, the introduction of a speed component to the RAPM test may change the operational definition of *g* to have a greater association with both content and process abilities than the original RAPM.

4.1. Remaining questions

One of the major limitations in the evaluation of data from speeded and standard versions of the RAPM, as noted earlier, is that even after several decades of existence, there are no alternate forms for the test. Having alternate forms would allow for the computation of an index of reliability that eliminates the potential contamination from temporary influences and lasting/specific influences. Alternate forms would also provide for a variety of other important considerations, such as the item-specific and test-general practice effects, along with an optimal assessment of transfer effects or intervention effects with the entire test. It is indeed a puzzle that such forms have not been developed and evaluated for these purposes.

Another limitation that became clear from our survey of the literature for the current meta-analysis is that numerous researchers fail to take account of appropriate measures of test reliability (i.e., not using internal consistency indicators), and make corrections for test reliability in their reporting of associations between RAPM measures and other measures of interest. Given that both shorter-than-standard measures and speeded versions (or some combination of the two) result in lower RAPM reliabilities, ceteris paribus, decisions to use such measures in any investigation are likely to underestimate convergent construct validity of their measures and overestimate discriminant validity estimates.

A third limitation from the literature was that authors who used the 'non-standard' versions of the RAPM frequently failed to describe the specific conditions under which the RAPM was administered – most notably in terms of the instructions and practice provided prior to administration of the test. That is, in the standard version, examinees are provided with interactive and worked instructions for the first two items of Set I of the RAPM, and then 5 min of experience in completing the remaining 10 items from Set I; all prior to completion of the 36 Set II

items. For the articles we reviewed, a few authors reported an abbreviated set of instructions and two or three practice items, but generally did not report a standard set of instructions and a complete administration of Set I. It remains to be demonstrated what effect, if any, such non-standard pre-test experience has on the Set II test scores and their correlations with other trait measures (though see Bors & Stokes, 1998, for one example).

A final limitation associated with this study involves the speed constraints imposed during the administration of ability tests other than RAPM. If, for example, tests that measure VZ that are highly speeded are administered alongside versions of RAPM that are highly speeded, the difference in estimated true score correlations between VZ and RAPM scores across RAPM speed conditions may be the result of the speed constraints imposed during the administration of the VZ test rather than RAPM measure. However, we should note that this is a somewhat complex issue – because it mainly applies to only 'some' ability assessments. For example, traditional assessments of Perceptual Speed tests are inherently highly speeded, regardless of the imposed limit for the entire test, and they are typically designed to provide far more potential items to be answered than any examinee could complete within the time available (see for example, Ackerman & Cianciolo, 2000). Psychomotor ability tests can be divided into two main categories – one that involves highly-speeded responses similar to the PS tests, and the other with 'steadiness' (e.g., Salvendy's One-Hole Test) or time-on-target (e.g., Rotary Pursuit). None of the studies included in the meta-analysis included data from either steadiness-type tests or time-on-target type tests. In addition, many memory tests have a completely different format, which does not encounter the speededness issue that is under discussion in this manuscript. That is, in a paired-associates type of test (e.g., the ETS first and last name test), the examinee is provided with a page of names to study [so that the first place speededness could be encountered is the total amount of study time, and then a page with the initial word prompts, for the examinee to respond with the appropriate second-word responses]. In both instances, however: (a) few, if any, implementations increase the speededness of the response phase to put significant speed-pressure on the examinee, and (b) there was insufficient information to determine how 'relatively' speeded the study phase was for the individual tests (per item) across studies.

In order to explore the extent to which the speediness of other ability tests may have confounded the difference in estimated true score correlations between such ability tests and RAPM under different RAPM administrations, we investigated the difference in speed constraints imposed during the administration of tests of Spatial Visualization and Closure – two of the ability categories to which this issue is immediately relevant. For each of the two abilities, we computed ANOVAs to determine whether or not there is a significant difference in the speed constraints (seconds per/item) imposed during administration of Spatial Visualization/Closure tests that were administered along with RAPM measures administered under standard, faster than standard, and slower than standard conditions. For each category, there was not a statistically significant difference in seconds/per item imposed across RAPM speed constraint categories ($F_{VZ} = 2.362$, p = ns; $F_{Closure} = 1.245$, p = ns). Therefore, in this subset of abilities, we did not find evidence to suggest that the difference in estimated true score correlations across the speededness in which RAPM is administered is confounded by the speededness by which other ability tests are administered.

5. Conclusions

It has been well over 100 years since the introduction of the first modern intelligence test by Binet and Simon (1905/1961). Quite a lot is known about the validity, reliability, and various characteristics of multiple tests and versions inspired by the Binet-Simon scales. For example, a Google Scholar search indicates roughly 76,200 items that reference the Stanford-Binet translations and revisions of the Binet-Simon Scales. A search of Wechsler's Adult Intelligence (first introduced as the Wechsler-Bellevue; Wechsler, 1939) indicated 190,000 items on Google Scholar. A search for Raven's Progressive Matrices resulted in 37,900 items. Yet, as indicated by the current meta-analysis, even though the RAPM has been largely unchanged in content or items since 1962 (or really, since 1947; given that the current version is mainly a subset of the items introduced in that version), there is still a substantial debate about *what* is actually measured by the RAPM.

The popularity of the RAPM for numerous investigations is likely at least partially attributable to Spearman's endorsement of the test as a 'perhaps the best of all non-verbal tests of G' (Spearman & Jones, 1950, p. 70), and at least partially attributable to the fact that the test is made up of a set of items that are, at least on the surface (and demonstrated in internal consistency estimates), homogeneous in content, appear to be minimally dependent on prior knowledge or verbal abilities and skills. For some investigators, the main disadvantage of RAPM is that, when administered with procedures specified in the manual (a 5-min Set I practice sequence with 2 worked examples and 10 practice items that precedes a 40 min administration of the 36 Set II items), the test does not fit easily into a laboratory or on-line study where there are significant time constraints (e.g., when undergraduate volunteers are tested, or examinees are given financial compensation for their study participation). The introduction of abbreviated versions of the RAPM, and the associated common reductions in administration time appear to ameliorate some of these logistical limitations.

However, as shown in the current meta-analysis, decreasing the number of items on the test and/or decreasing the time/item for the test, results in notable changes to the correlations between RAPM scores and other abilities, most notably Vz and PS-Memory. Based on these differences, the inference from these results is that the processes/abilities that are most influential in determining individual differences in performance on the RAPM are different when there are changes in the administration format. In addition, these differences are exaggerated to a degree for short versions of the RAPM, because of the reduction of test reliability associated with reducing a test to 1/2 or 1/3 of the original test length. It is useful to reiterate that, in addition to test length issues, various investigators over the last 70 years have expressed other concerns about the nature of the test construction, the unitary nature of the underlying test responses, and ultimately the relationship between the test and *Gf* or *g* (Burke, 1958; Gignac, 2015).

5.1. Assessing the RAPM and intellectual abilities

Although there were notable differences between correlations among the RAPM and other abilities, depending on (a) the administration time/item of the test and (b) whether the number of items administered on the test were substantially limited or not, many of the corresponding confidence intervals were overlapping, limited by the diversity of correlations reported in the literature. Because there are relatively few studies in many of the categories of abilities, it is not possible to definitively delineate the reasons for the observed pattern of effects. Moreover, as noted earlier, because of frequently unstated testing conditions reported in the articles (e.g., instructions, number of practice items), it is possible that some of the lack of decisive differences between correlations may have been the result of study differences that could not be ascertained.

Nonetheless, there are several conclusions that we believe can be drawn from the meta-analysis, in conjunction with prior literature on the Raven's Progressive Matrices test, as follows:

1. RAPM scores, especially when administered in a faster-than-standard testing format, are as highly related, or more highly related to spatial abilities (Vz, but also SR) as they are to omnibus IQ measures or tests/composites that represent *Gf*.
2. Similarly, faster-than-standard RAPM administrations are associated with significant and substantially larger correlations with tests of PS-Memory than are RAPM administrations with standard time limits.

3. Across-the-board of test administration time constraints, RAPM scores are as highly or more highly correlated with Short-Term Memory as they are with Working Memory (consistent with the empirical results of Martínez et al., 2011).
4. Longer (more items) versions of the RAPM tend to be more highly correlated with measures of a variety of different abilities, a pattern that is more pronounced if the correlations are not adjusted/corrected for the lower reliabilities of the shorter versions, compared to the longer versions of the RAPM. Ceteris paribus, the shorter versions appear to show less of an association with other abilities, including Vz, SR, IQ, and *Gf* than the longer versions, but a substantial portion of these differences can be attributed to reliability differences, except for Vz, which is more highly correlated with longer versions of the RAPM. Such results may indicate salient differences in the solution strategies by examinees who are engaged in the short versions of the RAPM. (Again, whether this is primarily a function of the *length* differences for the RAPM, the differences in instructions and pre-test practice, or both, are unknown potential influences.)
5. Finally, although there was an insufficient number of studies of criterion-related validity associated with the RAPM, the results of the current meta-analysis indicating differences in reliability and construct validity for the short/time-limited tests when compared to longer/less time constrained administrations suggests that resolving concerns about the criterion-related validity of the Raven's Progressive Matrices test may hinge, at least partly on the administration format. We note that Vernon and Parry's (1949) conclusions from large-scale criterion validity evaluation of the Raven's Progressive Matrices: "… rather poor reliability, and its susceptibility to non-intellectual influences" (p. 235) may have at least partly been a function of administering the test with a 20-min time limit. Consideration of the effects of short and/or speeded versions of the Raven, along with the Brunswik Symmetry perspective for the predictor-criterion space, may indeed suggest a scenario where the conditions of testing can be better aligned with the criterion of performance in the real-world.

We are hopeful that these results will encourage researchers to take heed of Boring's observation that there is "no essential difference between a mental test and a scientific psychological experiment", and that the conditions of testing should be a *substantive* consideration when using the RAPM or other similar tests, especially as a reference for *g* or *Gf*. The fact that a performance on short-form version or a speeded version of a test correlates strongly with performance on a standard version of the same test does not necessarily mean the tests are assessing the same construct or combinations of constructs. Our results indicate that alterations in test speededness or test length may increase the dependence of performance on some abilities and decrease the emphasis placed on others. This is not to say that short-form or speeded tests are always problematic. In many cases, such alterations may be appropriate for a given research question. However, selections of non-standard test versions seem to have been made without giving adequate attention to these considerations.

Funding

This research was partially supported by a grant from the U.S. Army Research Institute for the Social and Behavioral Sciences (W911NF-19-1-0400). The view, opinions, and/or findings contain in this report are those of the authors and shall not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documents.

Appendix A. Appendix

Representative tests for each ability category.

Ability Category	Representative Tests	Example Sources
Spatial Vz	WAIS Block Design	Wechsler, (2008)
	ETS Paper Folding	Ekstrom et al., (1976)
SR	Horn's Mental Rotation Scale	Horn, (1983)
	PMA-Spatial Rotation Test	Thurstone, (1962)
Closure	Group Embedded Figures Test	Witkin et al., (1971)
	ETS Gestalt Completion	Ekstrom et al., (1976)
Verbal ability	Mill Hill Vocabulary	Raven, (1958)
	WAIS-Verbal	Wechsler, (2008)
Verbal/fluency	Completion Test	Ebbinghaus, (1897)
	Remote Associates Test	Medick & Mednick, (1967)
Math	WAIS-Arithmetic	Wechsler, (2008)
	SAT Math	Shaw, (2015)
Gf	Cattell Culture Fair Test	Cattell, (1940)
	MAB-Performance Scale	Jackson, (1985)
Attention	Antisaccade task	Hallett & Adams, (1980)
	Flanker task	Eriksen & Eriksen, (1974)
Working memory	Operation span	Conway et al., (2005)
	N-back task	Jaeggi et al., (2003)
Short term memory	Digit span	Kane et al. (2004)
	Letter span	Kane et al. (2004)
Memory	ETS Shape Memory	Ekstrom et al., (1976)
	WJII: Picture Recognition	Woodcock et al., (2001)
Learning	Paired-Associate Learning	Calkins, (1894)
	ETS-Picture Number	Ekstrom et al., (1976)
Gc	MAB-Information	Jackson, (1985)
	WAIS-R-Information	Wechsler, (2008)
Knowledge	ASVAB-General Science	U.S. Department of Defense, (1984)
	Education Progress Test-History & Literature	Ravitch & Finn, (1987)
PS-Scanning	WJIII-Number Comparison Task	Woodcock et al., (2001)
	ETS-Identical Pictures	Ekstrom et al., (1976)
PS-Memory	WAIS-Digit Symbol	Wechsler, (2008)
	WAIS-Symbol Search	Wechsler, (2008)
PS-Pattern Recognition	ETS-Findings A's	Ekstrom et al., (1976)

(*continued on next page*)

(*continued*)

Ability Category	Representative Tests	Example Sources
	Zahlen-Verbindungen Test	Vernon (1993)
Psychomotor	Circle Tapping	Fleishman, (1954)
	Go/no go task	Donders, (1868)

Spatial Vz = Spatial Visualization, WAIS = Wechsler Adult Intelligence Scale, ETS = Educational Testing Services, SR = Speeded Rotation, PMA = Primary Mental Abilities, Gf = Fluid Intelligence, MAB = Multidimensional Aptitude Battery, WJIII = Woodcock-Johnson Third Edition, Gc = Crystallized Intelligence, WAIS-R = Wechsler Adult Intelligence Scale – Revised, ASVAB = Armed Services Vocational Aptitude Battery, PS = Perceptual Speed.

References

*Ackerman, P. L. (1986). Individual differences in information processing: An investigation of intellectual abilities and task performance during practice. *Intelligence, 10*(2), 101–139.

*Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117*(3), 288–318.

*Ackerman, P. L. (1990). A correlational analysis of skill specificity: Learning, abilities, and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(5), 883–901.

*Ackerman, P. L. (1992). Predicting individual differences in complex skill acquisition: Dynamics of ability determinants. *Journal of Applied Psychology, 77*(5), 598–614.

*Ackerman, P. L. (2000). Domain-specific knowledge as the" dark matter" of adult intelligence: Gf/Gc, personality and interest correlates. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 55*(2), P69–P84.

*Ackerman, P. L., & Beier, M. E. (2007). Further explorations of perceptual speed abilities in the context of assessment methods, cognitive abilities, and individual differences during skill acquisition. *Journal of Experimental Psychology: Applied, 13*(4), 249–272.

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131*(1), 30–60.

*Ackerman, P. L., Beier, M. E., & Boyle, M. D. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General, 131*(4), 567–589.

Ackerman, P. L., & Cianciolo, A. T. (2000). Cognitive, perceptual-speed, and psychomotor determinants of individual differences during skill acquisition. *Journal of Experimental Psychology: Applied, 6*(4), 259–290.

*Ackerman, P. L., & Kanfer, R. (1993). Integrating laboratory and field study for improving selection: Development of a battery for predicting air traffic controller success. *Journal of Applied Psychology, 78*(3), 413–432.

Ackerman, P. L., & Lohman, D. F. (2006). Individual differences in cognitive functions. In P. A. Alexander, & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 139–161). Lawrence Erlbaum Associates Publishers.

*Ackerman, P. L., & Wolman, S. D. (2007). Determinants and validity of self-estimates of abilities and self-concept measures. *Journal of Experimental Psychology: Applied, 13*(2), 57–78.

*Alberts, P. P. H. (2007). *The predictive validity of a selection battery for university bridging students in a public sector organisation* (Doctoral dissertation. North-West University.

*Allan, J. N. (2018). Numeracy vs. In *Intelligence: A model of the relationship between cognitive abilities and decision making*. University of Oklahoma. Doctoral dissertation.

Arai, T. (1912). *Mental fatigue (No. 54). Teachers College*. Columbia University.

*Arthur, W., Jr., & Day, D. V. (1991). Examination of the construct validity of alternative measures of field dependence/independence. *Perceptual and Motor Skills, 72*(3), 851–859.

Arthur, W., Jr., & Day, D. V. (1994). Development of a short form for the Raven advanced progressive matrices test. *Educational and Psychological Measurement, 54*(2), 394–403.

Arthur Jr., W., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the Raven Advanced Progressive Matrices Test. *Journal of Psychoeducational Assessment, 17*(4), 354–361.

*Babcock, R. L. (1994). Analysis of adult age differences on the Raven's advanced progressive matrices test. *Psychology and Aging, 9*(2), 303–314.

*Babcock, R. L., & Laguna, K. D. (1996). An examination of the adult age-related differences on the Raven's advanced progressive matrices: A structural equations approach. *Aging, Neuropsychology, and Cognition, 3*(3), 187–200.

*Baghaei, P., Khoshdel-Niyat, F., & Tabatabaee-Yazdi, M. (2017). *The Persian adaptation of Baddeley's 3-min grammatical reasoning test* (pp. 30–35). Psicologia: Reflexão e Crítica.

*Batey, M., Furnham, A., & Safiullina, X. (2010). Intelligence, general knowledge and personality as predictors of creativity. *Learning and Individual Differences, 20*(5), 532–535.

Binet, A., & Simon, T. (1905/1961). New methods for the diagnosis of the intellectual level of subnormals. Trans.. In E. S. Kite, J. J. Jenkins, & D. G. Paterson (Eds.), *Studies of individual differences: The search for intelligence* (pp. 90–96). New York, NY: Appleton-Century-Crofts (Reprinted from 1905, L'Année Psychologique, 11, 191–244, 1905.)

*Birney, D. P., Beckmann, J. F., Beckmann, N., & Double, K. S. (2017). Beyond the intellect: Complexity and learning trajectories in Raven's progressive matrices depend on self-regulatory processes and conative dispositions. *Intelligence, 61*, 63–77.

Borenstein, M., Hedges, L. V., & Rothstein, H. R. (2007). *Meta-analysis. Fixed effects vs random effects*. Englwood, NJ: Biostat.

Bors, D. A., & Stokes, T. L. (1998). Raven's advanced progressive matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement, 58*(3), 382–398.

Bors, D. A., & Vigneau, F. (2003). The effect of practice on Raven's advanced progressive matrices. *Learning and Individual Differences, 13*(4), 291–312.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322.

*Bruza, B., Welsh, M. B., & Navarro, D. J. (2008). Does memory mediate susceptibility to cognitive biases? Implications of the decision-by-sampling theory. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1498–1503). Austin, TX: Cognitive Science Society.

*Buckley, J., Seery, N., Canty, D., & Gumaelius, L. (2018). Visualization, inductive reasoning, and memory span as components of fluid intelligence: Implications for technology education. *International Journal of Educational Research, 90*, 64–77.

*Burgoyne, A. P., Hambrick, D. Z., & Altmann, E. M. (2019). Is working memory capacity a causal factor in fluid intelligence? *Psychonomic Bulletin & Review, 26*(4), 1333–1339.

*Burgoyne, A. P., Harris, L. J., & Hambrick, D. Z. (2019). Predicting piano skill acquisition in beginners: The role of general intelligence, music aptitude, and mindset. *Intelligence, 76*, Article 101383.

Burke, H. R. (1958). Raven's progressive matrices: A review and critical evaluation. *The Journal of Genetic Psychology, 93*(2), 199–228.

Calkins, M. W. (1894). Experimental. *Psychological Review, 1*(3), 327–329.

Carroll, J. B. (1982). The measurement of intelligence. In R. X. Sternberg (Ed.), *Handbook of human intelligence*. Cambridge, England: Cambridge University Press.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

Cattell, R. B. (1940). A culture free intelligence test. I. *Journal of Educational Psychology, 31*(3), 161–179.

*Chiesi, F., Ciancaleoni, M., Galli, S., Morsanyi, K., & Primi, C. (2012). Item response theory analysis and differential item functioning across age, gender and country of a short form of the advanced progressive matrices. *Learning and Individual Differences, 22*(3), 390–396.

*Choi, J., & L'Hirondelle, N. (2005). Object location memory: A direct test of the verbal memory hypothesis. *Learning and Individual Differences, 15*(3), 237–245.

*Chow, M. A. (2017). *Decoding 3 categories of conventional wisdom in pattern similarity analyses*. Doctoral dissertation. Princeton University.

*Chuderski, A. (2013). When are fluid intelligence and working memory isomorphic and when are they not? *Intelligence, 41*(4), 244–262.

*Cockcroft, K., & Israel, N. (2011). The Raven's advanced progressive matrices: A comparison of relationships with verbal ability tests. *South Africa Journal of Psychology, 41*(3), 363–372.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

*Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision making, 7*(1), 25–47.

*Colom, R., Escorial, S., & Rebollo, I. (2004). Sex differences on the progressive matrices are influenced by sex differences on spatial ability. *Personality and Individual Differences, 37*(6), 1289–1293.

*Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. C. (2004). Working memory is (almost) perfectly predicted by g. *Intelligence, 32*(3), 277–296.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*(5), 769–786.

*Coyle, T. R., & Pillow, D. R. (2008). SAT and ACT predict college GPA after removing g. *Intelligence, 36*(6), 719–729.

*Culbertson, S. S., Huffcutt, A. I., & Goebl, A. P. (2013). Introduction and empirical assessment of executive functioning as a predictor of job performance. *PsyCh Journal, 2*(2), 75–85.

*Dang, C. P., Braeken, J., Ferrer, E., & Liu, C. (2012). Unitary or non-unitary nature of working memory? Evidence from its relation to general fluid and crystallized intelligence. *Intelligence, 40*(5), 499–508.

*Darowski, E. S., Helder, E., Zacks, R. T., Hasher, L., & Hambrick, D. Z. (2008). Age-related differences in cognition: The role of distraction control. *Neuropsychology, 22*(5), 638–644.

*De Simoni, C., & von Bastian, C. C. (2018). Working memory updating and binding training: Bayesian evidence supporting the absence of transfer. *Journal of Experimental Psychology: General, 147*(6), 829–858.

*DeYoung, C. G., Peterson, J. B., & Higgins, D. M. (2005). Sources of openness/intellect: Cognitive and neuropsychological correlates of the fifth factor of personality. *Journal of Personality, 73*(4), 825–858.
*Dodonova, Y. A., & Dodonov, Y. S. (2012). Processing speed and intelligence as predictors of school achievement: Mediation or unique contribution? *Intelligence, 40*(2), 163–171.
Ebbinghaus, H. (1897). UÈ ber eine neue Methode zur PruÈfung geistiger FaÈhigkeiten und ihre Anwendung bei Schulkindem (On a new method for testing mental abilities and its use with school children). *Zeitschrift fuÈr Psychologie und Psysiologie der Sinnesorgane, 13*, 401–459 (translated by Wilhelm, 1999).
*Edwards, B. D. (2004). *An examination of factors contributing to a reduction in race-based subgroup differences on a constructed response paper-and-pencil test of achievement*. doctoral dissertation. Texas A&M University.
Ekstrom, R. B., French, J. W., Harman, H., & Derman, D. (1976). *Kit of Factor-Referenced Cognitive Tests (revised edition)*. Princeton, NJ: Educational Testing Service.
*Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 380.
Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*(1), 19–23.
Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16*(1), 143–149.
*Ettinger, U., & Corr, P. J. (2001). The frequency accrual speed test (FAST): Psychometric intelligence and personality correlates. *European Journal of Personality, 15*(2), 143–152.
Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement, 9*(2), 122–131.
*Felez-Nobrega, M., Foster, J. L., Puig-Ribera, A., Draheim, C., & Hillman, C. H. (2018). Measuring working memory in the Spanish population: Validation of a multiple shortened complex span task. *Psychological Assessment, 30*(2), 274.
Fleishman, E. A. (1954). Dimensional analysis of psychomotor abilities. *Journal of Experimental Psychology, 48*(6), 437–454.
Fry, A. F., & Hale, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science, 7*(4), 237–241.
*Furlan, S. (2011). *Developmental and individual differences in the ratio-Bias phenomenon with and without time pressure*. doctoral dissertation. Universita Delgi Studi Di Padova.
*Furlan, S., Agnoli, F., & Reyna, V. F. (2016). Intuition and analytic processes in probabilistic reasoning: The role of time pressure. *Learning and Individual Differences, 45*, 1–10.
Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for *g* factor theory and the brief measurement of *g*. *Intelligence, 52*, 71–79.
*Graham, J. D. B. (2011). *Elements of human effectiveness: Intelligences, traits, and abilities that lead to success and fulfillment in life*. doctoral dissertation. Irvine: University of California.
*Greengross, G., & Miller, G. (2011). Humor ability reveals intelligence, predicts mating success, and is higher in males. *Intelligence, 39*(4), 188–192.
*Griffin, B., Carless, S., & Wilson, I. (2013). The undergraduate medical and health sciences admissions test: What is it measuring? *Medical Teacher, 35*(9), 727–730.
*Grounds, M. (2016). *Communicating weather uncertainty: An individual differences approach*. doctoral dissertation. University of Washington.
*Gutierrez, J. C., Holladay, S. D., Arzi, B., Gomez, M., Pollard, R., Youngblood, P., & Srivastava, S. (2018). Entry-level spatial and general non-verbal reasoning: Can these abilities be used as a predictor for anatomy performance in veterinary medical students? *Frontiers in Veterinary Science, 5*, 226.
*Guye, S., & von Bastian, C. C. (2017). Working memory training in older adults: Bayesian evidence supporting the absence of transfer. *Psychology and Aging, 32*(8), 732–746.
*Haavisto, M. L., & Lehto, J. E. (2005). Fluid/spatial and crystallized intelligence in relation to domain-specific working memory: A latent-variable approach. *Learning and Individual Differences, 15*(1), 1–21.
*Haier, R. J., Siegel, B., Tang, C., Abel, L., & Buchsbaum, M. S. (1992). Intelligence and changes in regional cerebral glucose metabolic rate following learning. *Intelligence, 16*(3–4), 415–426.
Hallett, P. E., & Adams, B. D. (1980). The predictability of saccadic latency in a novel voluntary oculomotor task. *Vision Research, 20*(4), 329–339.
Hamel, R., & Schmittmann, V. D. (2006). The 20-minute version as a predictor of the Raven advanced progressive matrices test. *Educational and Psychological Measurement, 66*(6), 1039–1046.
*Hancock, D. G. (2017). *An exploratory look at grit's differential relationships with facets of conscientiousness*. Doctoral dissertation. The University of Texas at San Antonio.
*Hannon, B. (2016). General and non-general intelligence factors simultaneously influence SAT, SAT-V, and SAT-M performance. *Intelligence, 59*, 51–63.
*Hannon, B., & Daneman, M. (2014). Revisiting the construct of "relational integration" and its role in accounting for general intelligence: The importance of knowledge integration. *Intelligence, 47*, 175–187.
*Hardmeier, D., & Schwaninger, A. (2008). Visual cognition abilities in x-ray screening. In *Proceedings of the Third International Conference on Research in Air Transportation, June, 2008* (pp. 311–316).
*Hicks, K. L., Foster, J. L., & Engle, R. W. (2016). Measuring working memory capacity on the web with the online working memory lab (the OWL). *Journal of Applied Research in Memory and Cognition, 5*(4), 478–489.
Horn, W. (1983). *LPS: Leistungs-Prüf-System*. Göttingen, Germany: Hogrefe.
*Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience, 31*(1), 80–93.
Hunt, E., Frost, N., & Lunneborg, C. (1973). Individual differences in cognition: A new approach to intelligence. *Psychology of Learning and Motivation, 7*, 87–122.
*Hunt, E., Pellegrino, J. W., Frick, R. W., Farr, S. A., & Alderton, D. (1988). The ability to reason about movement in the visual field. *Intelligence, 12*(1), 77–100.
Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.
Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8*, 275–292.
*Israel, N. (2006). *Raven's advanced progressive matrices within a south African context*. Doctoral dissertation. University of Witwatersrabdm Johannesburg.
Jackson, D. N. (1985). *Multidimensional Aptitude Battery*. London, Ontario, Canada: Sigma Assessment Systems.
Jaeggi, S. M., Seewer, R., Nirkko, A. C., Eckstein, D., Schroth, G., Groner, R., & Gutbrod, K (2003). Does excessive memory load attenuate activation in the prefrontal cortex? Load-dependent processing in single and dual tasks: Functional magnetic resonance imaging study. *NeuroImage, 19*(2), 210–225.
*Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y. F., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning—Implications for training and transfer. *Intelligence, 38*(6), 625–635.
*Jarosz, A. F., & Wiley, J. (2012). Why does working memory capacity predict RAPM performance? A possible role of distraction. *Intelligence, 40*(5), 427–438.
Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.
*Kaesler, M., Welsh, M., & Semmler, C. (2016). Predicting overprecision in range estimation. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society (CogSci, 2016), June, 2016* (pp. 502–507).
*Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and Visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133*(2), 189–217.
*Kaufman, S. B., DeYoung, C. G., Gray, J. R., Brown, J., & Mackintosh, N. (2009). Associative learning predicts intelligence above and beyond working memory and processing speed. *Intelligence, 37*(4), 374–382.
*Kaufman, S. B., DeYoung, C. G., Reis, D. L., & Gray, J. R. (2011). General intelligence predicts reasoning ability even for evolutionarily familiar content. *Intelligence, 39*(5), 311–322.
*Kock, F. D., & Schlechter, A. (2009). Fluid intelligence and spatial reasoning as predictors of pilot training performance in the south African air force (SAAF). *SA Journal of Industrial Psychology, 35*(1), 31–38.
*Koenig, K. A., Frey, M. C., & Detterman, D. K. (2008). ACT and general cognitive ability. *Intelligence, 36*(2), 153–160.
*Kpolovie, P. J., & Emekene, C. O. (2016). Item response theory validation of advanced progressive matrices in Nigeria. *British Journal of Psychology Research, 4*(1), 1–32.
*Kranzler, J. H., & Jensen, A. R. (1991). The nature of psychometric g: Unitary process or a number of independent processes? *Intelligence, 15*(4), 397–422.
*Kulikowski, K., & Orzechowski, J. (2019). Working memory and fluid intelligence as predictors of work engagement—Testing preliminary models. *Applied Cognitive Psychology, 33*(4), 596–616.
Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14*(4), 389–433.
*Lee, C. S., & Therriault, D. J. (2013). The cognitive underpinnings of creative thought: A latent variable analysis exploring the roles of intelligence and working memory in three creative thinking processes. *Intelligence, 41*(5), 306–320.
*Li, S., Ren, X., Schweizer, K., Brinthaupt, T. M., & Wang, T. (2021). Executive functions as predictors of critical thinking: Behavioral and neural evidence. *Learning and Instruction, 71*, Article 101376.
*Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making, 25*(4), 361–381.
*Lilienthal, L., Tamez, E., Myerson, J., & Hale, S. (2013). Predicting performance on the Raven's matrices: The roles of associative learning and retrieval efficiency. *Journal of Cognitive Psychology, 25*(6), 704–716.
*Lin, W.-L., Hsu, K.-Y., Chen, H.-C., & Wang, J.-W. (2012). The relations of gender and personality traits on different creativities: A dual-process theory account. *Psychology of Aesthetics, Creativity, and the Arts, 6*(2), 112–123.
Lord, F. M. (1956). A study of speed factors in tests and academic grades. *Psychometrika, 21*(1), 31–50.
*Mackintosh, N. J., & Bennett, E. S. (2003). The fractionation of working memory maps onto different components of intelligence. *Intelligence, 31*(6), 519–531.
*Mackintosh, N. J., & Bennett, E. S. (2005). What do Raven's matrices measure? An analysis in terms of sex differences. *Intelligence, 33*(6), 663–674.
*Martin, J., Mashburn, C. A., & Engle, R. W. (2020). Improving the validity of the armed service vocational aptitude battery with measures of attention control. *Journal of Applied Research in Memory and Cognition, 9*(3), 323–335.
*Martinez, D. (2019). Immediate and long-term memory and their relation to crystallized and fluid intelligence. *Intelligence, 76*, Article 101382.
Martínez, K., Burgaleta, M., Román, F. J., Escorial, S., Shih, P. C., Quiroga, M.Á., & Colom, R. (2011). Can fluid intelligence be reduced to 'simple' short-term storage? *Intelligence, 39*(6), 473–480.
*McCrory, C., & Cooper, C. (2007). Overlap between visual inspection time tasks and general intelligence. *Learning and Individual Differences, 17*(2), 187–192.
*McPherson, J., & Burns, N. R. (2007). Gs invaders: Assessing a computer game-like test of processing speed. *Behavior Research Methods, 39*(4), 876–883.

*McPherson, J., & Burns, N. R. (2008). Assessing the validity of computer-game-like tests of processing speed and working memory. *Behavior Research Methods, 40*(4), 969–981.
*McRorie, M., & Cooper, C. (2003). Neural transmission and general mental ability. *Learning and Individual Differences, 13*(4), 335–338.
*McRorie, M., & Cooper, C. (2004a). Psychomotor movement and IQ. *Personality and Individual Differences, 37*(3), 523–531.
*McRorie, M., & Cooper, C. (2004b). Synaptic transmission correlates of general mental ability. *Intelligence, 32*(3), 263–275.
Mednick, S. A., & Mednick, M. T. (1967). *Examiner's manual: Remote Associates Test*. Boston: Houghton Mifflin.
*Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., … Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied, 21*(1), 1–14.
*Miroshnik, K. G., & Shcherbakova, O. V. (2019). The proportion and creativity of "old" and "new" ideas: Are they related to fluid intelligence? *Intelligence, 76*, Article 101384.
Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), Article e1000097.
*Morsanyi, K., Handley, S. J., & Serpell, S. (2013). Making heads or tails of probability: An experiment with random generators. *British Journal of Educational Psychology, 83*(3), 379–395.
*Morsanyi, K., O'Mahony, E., & McCormack, T. (2017). Number comparison and number ordering as predictors of arithmetic performance in adults: Exploring the link between the two skills, and investigating the question of domain-specificity. *Quarterly Journal of Experimental Psychology, 70*(12), 2497–2517.
*Naber, A. M. (2015). *Increased retest scores on cognitive tests: Learning or memory effects?*. Doctoral dissertation. Texas A&M University.
*Neubauer, A. C., & Bucik, V. (1996). The mental speed—IQ relationship: Unitary or modular? *Intelligence, 22*(1), 23–48.
*Oswald, F. L., McAbee, S. T., Redick, T. S., & Hambrick, D. Z. (2015). The development of a short domain-general measure of working memory capacity. *Behavior Research Methods, 47*(4), 1343–1355.
*Pahor, A., Stavropoulos, T., Jaeggi, S. M., & Seitz, A. R. (2019). Validation of a matrix reasoning task for mobile devices. *Behavior Research Methods, 51*(5), 2256–2267.
*Park, I., & Cho, S. (2019). The influence of number line estimation precision and numeracy on risky financial decision making. *International Journal of Psychology, 54*(4), 530–538.
*Paul, S. M. (1986). The advanced Raven's progressive matrices: Normative data for an American university population and an examination of the relationship with Spearman's g. *The Journal of Experimental Education, 54*(2), 95–100.
Penrose, L. S., & Raven, J. C. (1936). A new series of perceptual tests: Preliminary communication. *British Journal of Medical Psychology, 16*(2), 97–104.
Donders, F. C. (1868). Die schnelligkeit psychischer processe: Erster artikel. *Archiv für Anatomie, Physiologie und wissenschaftliche Medicin*, 657–681.
*du Pont, A., Karbin, Z., Rhee, S. H., Corley, R. P., Hewitt, J. K., & Friedman, N. P. (2020). Differential associations between rumination and intelligence subtypes. *Intelligence, 78*, Article 101420.
*Prasad, J. (2014). *Connecting the dots between N-back, operation span, and Raven's progressive matrices*. Doctoral dissertation. Wake Forest University.
Quereshi, M. Y. (1960). Mental test performance as a function of payoff conditions, item difficulty, and degree of speeding. *Journal of Applied Psychology, 44*(2), 65–77.
Quintana, D. S. (2015). From pre-registration to publication: A non-technical primer for conducting a meta-analysis to synthesize correlational data. *Frontiers in Psychology, 6*, 1549.
R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Raven, J., Raven, J. C., & Court, J. H. (1998). *Advanced progressive matrices: 1998 edition: Raven manual: Section 4*. Oxford: Oxford Psychologists Press.
Raven, J. C. (1965). *Guide to using the Mill Hill Vocabulary Scale with the Progressive Matrices Scale*. London: H.K. Lewis.
Raven, J. C., Raven, J., & Court, J. H. (1962). *Advanced progressive matrices set II*. Oxford, England: Oxford Psychologists Press.
Ravitch, D., & Finne, C. E. (1987). *What do our 17-year-olds know?*. New York: Harper & Row.
*Ren, X., Schweizer, K., Wang, T., Chu, P., & Gong, Q. (2017). On the relationship between executive functions of working memory and components derived from fluid intelligence measures. *Acta Psychologica, 180*, 79–87.
*Ren, X., Tong, Y., Peng, P., & Wang, T. (2020). Critical thinking predicts academic performance beyond general cognitive ability: Evidence from adults and children. *Intelligence, 82*, Article 101487.
*Richmond, M. (2015). *G and non-g influences on GPA for Hispanics and whites: A structural equation modeling (SEM) approach to Spearman's law of diminishing returns (SLDOR)*. Doctoral Dissertation. The University of Texas at San Antonio.
*Rohde, T. E. (2008). *An examination of how visual perception abilities influence mathematics achievement*. Doctoral dissertation. Case Western Reserve University.
*Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence, 35*(1), 83–92.
*Saccuzzo, D. P., Craig, A. S., Johnson, N. E., & Larson, G. E. (1996). Gender differences in dynamic spatial abilities. *Personality and Individual Differences, 21*(4), 599–607.
Salthouse, T. A. (2014). Relations between running memory and fluid intelligence (Gf). *Intelligence, 43*, 1–7.
*Sanchez, C. A., Wiley, J., Miura, T. K., Colflesh, G. J., Ricks, T. R., Jensen, M. S., & Conway, A. R. (2010). Assessing working memory capacity in a non-native language. *Learning and Individual Differences, 20*(5), 488–493.
Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350–353.
*Schwarb, H. (2012). *Optimized cognitive training: Investigating the limits of brain training on generalized cognitive function*. Doctoral dissertation. Georgia Institute of Technology.
*Sefcek, J. A., & Figueredo, A. J. (2010). A life-history model of human fitness indicators. *Biodemography and Social Biology, 56*(1), 42–66.
*Seidler, T. (2014). *A beautiful mind: Examining the effects of emotional intelligence and physical attractiveness on employee evaluations*. Master's thesis. Western Kentucky University.
*Sevenants, A., Dieussaert, K., & Schaeken, W. (2013). Truth table task: Working memory load, latencies, and perceived relevance. *Journal of Cognitive Psychology, 25*(3), 339–364.
Shaw, E. J. (2015). *An SAT(r) Validity Primer.*. New York: College Board.
*Shelton, J. T., Elliott, E. M., Matthews, R. A., Hill, B. D., & Gouvier, W. M. (2010). The relationships of working memory, secondary memory, and general fluid intelligence: Working memory is special. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 813.
*Siebert, J. M. (2019). *Toward linguistically fair IQ screening: The multilingual vocabulary test*. Master's thesis. University of Cape Town.
Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120.
*Singh, K. A., Gignac, G. E., Brydges, C. R., & Ecker, U. K. (2018). Working memory capacity mediates the relationship between removal and fluid intelligence. *Journal of Memory and Language, 101*, 18–36.
*Skagerlund, K., Forsblad, M., Slovic, P., & Västfjäll, D. (2020). The Affect Heuristic and Risk Perception–Stability across elicitation methods and individual cognitive abilities. *Frontiers in Psychology, 11*.
Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72–101.
Spearman, C., & Jones, L. L. W. (1950). *Human ability: A continuation of "the abilities of man"*. London: Macmillan & Co.
*Srisang, P. (2017). *Influence of inferential skills on the reading comprehension ability of adult Thai (L1) and English (L2) students*. Doctoral Dissertation. University of Canterbury.
*Stanovich, K. E., & Cunningham, A. E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Memory & Cognition, 20*(1), 51–68.
*Tabatabaee-Yazdi, M., & Baghaei, P. (2018). Reading comprehension in English as a foreign language and some Cattell-horn-Carroll cognitive ability factors. *The Reading Matrix: An International Online Journal, 18*(1).
*Tabe, A. (2019). *Confidence as a predictor of academic success*. Undergraduate Thesis. University of Adelaide.
Terman, L. M. (1924). The mental test as a psychological method. *Psychological Review, 31*(2), 93–117.
*Teunisse, A. K., Case, T. I., Fitness, J., & Sweller, N. (2020). I should have known better: Development of a self-report measure of gullibility. *Personality and Social Psychology Bulletin, 46*(3), 408–423.
Thorndike, E. L. (1908). The effect of practice in the case of a purely intellectual function. *The American Journal of Psychology, 19*(3), 374–384.
Thorndike, E. L., Bregman, E. O., Cobb, M. V., Woodyard, E., & Inst of Educational Research Div of Psychology, Teachers Coll, Columbia U. (1926). *The measurement of intelligence*. New York: Teachers College Bureau of Publications.
Thorndike, R. L. (Ed.). (1947). *Army air forces aviation psychology program research reports: Research problems and techniques (report no. 3)*. Washington, DC: U.S. Government Printing Office.
Thurstone, T. G. (1962). *PMA [Primary Mental Abilities]*. Chicago: Science Research Associates.
U.S. Department of Defense. (1984). *Test Manual for the Armed Services Vocational Aptitude Battery*. North Chicago, IL: United States Military Entrance Processing Command.
Vernon, P. A. (1993). Der Zahlen-Verbindungs-test and other trail-making correlates of general intelligence. *Personality and Individual Differences, 14*(1), 35–40.
Vernon, P. E. (1947). Research on personnel selection in the Royal Navy and the British Army. *American Psychologist, 2*(2), 35–51.
Vernon, P. E., & Parry, J. B. (1949). *Personnel selection in the British forces*. London: University of London Press, Ltd.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.
*Villado, A. J., Randall, J. G., & Zimmer, C. U. (2016). The effect of method characteristics on retest score gains and criterion-related validity. *Journal of Business and Psychology, 31*(2), 233–248.
*Wang, Y. (2012). *The relationship between working memory and intelligence: Deconstructing the working memory task*. Doctoral dissertation. University of Florida.
*Waschl, N. A., Nettelbeck, T., & Burns, N. R. (2017). The role of visuospatial ability in the Raven's progressive matrices. *Journal of Individual Differences, 38*(4), 241–255.
Wechsler, D. (1939). *The measurement of adult intelligence*. Baltimore, MD: Williams & Wilkins.
Wechsler, D. (2008). *WAIS-IV administration and scoring manual*. San Antonio, TX: Psychological Corporation.
*Wei, W., Yuan, H., Chen, C., & Zhou, X. (2012). Cognitive correlates of performance in advanced mathematics. *British Journal of Educational Psychology, 82*(1), 157–181.
Wild, C. L., Durso, R., & Rubin, D. B. (1982). Effect of increased test-taking time on test scores by ethnic group, years out of school, and sex. *Journal of Educational Measurement, 19*(1), 19–28.

*Williams, B. A., & Pearlberg, S. L. (2006). Learning of three-term contingencies correlates with Raven scores, but not with measures of cognitive processing. *Intelligence, 34*(2), 177–191.
*Winman, A., Juslin, P., Lindskog, M., Nilsson, H., & Kerimi, N. (2014). The role of ANS acuity and numeracy for the calibration and the coherence of subjective probability judgments. *Frontiers in Psychology, 5*, 851.
Wittmann, W. W., & Süß, H.-M. (1999). Investigating the paths between working memory, intelligence, knowledge, and complex problem-solving performances via Brunswik symmetry. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 77–108). Washington, DC: American Psychological Association.
Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of achievement*. Itasca, IL: Riverside.
*Xie, Q. (2015). Intellectual styles: Their associations and their relationships to ability and personality. *Journal of Cognitive Education and Psychology, 14*(1), 63–76.
Yoakum, C. S., & Yerkes, R. M. (1920). *Army mental tests*. New York: H. Holt.
*Zajenkowski, M., Stolarski, M., Maciantowicz, O., Malesza, M., & Witowska, J. (2016). Time to be smart: Uncovering a complex interplay between intelligence and time perspectives. *Intelligence, 58*, 1–9.
*Zajenkowski, M., & Szymanik, J. (2013). MOST intelligent people are accurate and SOME fast people are intelligent: Intelligence, working memory, and semantic processing of quantifiers from a computational perspective. *Intelligence, 41*(5), 456–466.
*Zhu, B., Chen, C., Loftus, E. F., Lin, C., He, Q., Chen, C., … Dong, Q. (2010). Individual differences in false memory from misinformation: Cognitive factors. *Memory, 18*(5), 543–555.
*Zimowski, M. F., & Wothke, W. (1988). *The measurement of structural visualization: An evaluation of spatial and nonspatial sources of variation in the wiggly block and paper folding test scores* (Technical Report No. 1988–5).
*Zmigrod, L., & Zmigrod, S. (2016). On the temporal precision of thought: Individual differences in the multisensory temporal binding window predict performance on verbal and nonverbal problem solving tasks. *Multisensory Research, 29*(8), 679–701.