Intelligence 80 (2020) 101440

Contents lists available at ScienceDirect

Intelligence

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

Correspondence

A primer on assessing intelligence in laboratory studies

ABSTRACT

This paper is an attempt to provide a brief guide to major conceptual and statistical problems that are unique to the study of individual differences in intelligence and various intellectual abilities, in the context of laboratory experimental studies, and to suggest strategies to successfully navigate these problems. Such studies are generally designed so that the goal is to evaluate the relationships between individual differences in basic task performance or related markers on the one hand, and individual differences in intellectual abilities on the other hand. Issues discussed in this paper include: restriction-of-range in talent, method variance and facet theory; speed vs. power; regression to the mean; extreme-groups designs; difference scores; differences in correlations; significant vs. meaningful correlations; factor- pure tests; and criterion variables. A list of representative "do" and "don't" recommendations is provided to help guide the design and evaluation of laboratory studies.

1. Background

Ever since the groundbreaking work by Hunt and his colleagues (e.g., Hunt, Frost, and Lunneborg, 1973), numerous researchers have attempted to investigate the relationships between intellectual abilities on the one hand, and constructs from experimental cognitive psychology, on the other hand. However, there is a mismatch between the standard paradigm used in experimental psychology and the procedures that are optimal for investigating individual differences in intellectual abilities. Some of the difficulties may be due to the fundamentally different approaches taken by experimental and correlational (differential) psychology, as articulated by Cronbach (1957). In Cronbach's view, experimental psychologists are interested only in variation in behavior that they create through, for example, manipulation of experimental conditions and control conditions, or by parametrically varying a set of stimulus or response conditions. By contrast, correlational psychologists (including those who seek to assess intelligence) are interested in variation in behavior that exists in the world at large, which they largely cannot control or cannot even hope to control (e.g., genetics, education, socio-economic status).

This paper was written as a guide to the design of experiments that attempt to link laboratory-based assessment of basic cognitive/information processing constructs to individual differences in intellectual abilities. Many of the topics and problems reviewed here are not unique to such studies – for example, they also arise in studies that only consider relations among various intellectual abilities or other trait measures. However, the issues discussed here are notable because even though many of them are familiar to researchers who are concerned mainly with the study of individual differences, they are frequently not addressed adequately in experimental studies that include consideration of individual differences in intellectual abilities. Table 1 contains a list of "don't" and "do" recommendations for design and analysis for laboratory studies assessing intelligence. Each of these recommendations is discussed in turn.

2. Reliability and validity

The first difficulty in this area of research is that experimental and correlational orientations create a fundamental conflict. In the context of cognitive research, the experimental approach typically involves using stimuli that are either believed to be substantially "overlearned" by most or all study participants (e.g., letters, numbers, high-frequency words, simple polygons), or entirely novel stimuli (e.g., artificial grammars, random number sequences), such that either way, the expectation is that there will be little or no variance accounted for by participant familiarity with the material. To further minimize individual-difference variance in numerous variables, the typical laboratory experiment is based on participants who are highly homogeneous on age, educational attainment, and even socio-economic status (e.g., college/university freshmen). In contrast, because applications of intelligence assessments are in the realm of predicting real-world behaviors (e.g., academic success/failure) with examinees representing a wide range of backgrounds, the kinds of test (stimulus) materials on intelligence assessments also have a wide range of familiarity, ranging from well-learned to highly novel, but also many items somewhere in between familiar and novel, such as vocabulary, general information, math word problems, comprehension, and so on.

The second difficulty is that there is a frequent lack of understanding among experimental cognitive researchers about both reliability and validity concepts. Issues of reliability and validity are core concepts for assessment of individual differences in intellectual abilities, and they represent a central problem for choices to be made in the use of particular measures in the study of intelligence. As elementary textbooks on psychological assessment repeatedly assert (e.g., Anastasi & Urbina, 1997), individual tests do not have *inherent* reliability or validity; both are multifaceted concepts that are dependent on the population under investigation, the conditions of testing, *and* the inferences that the researcher intends to make.

2.1. Reliability

1. Don't disregard inadequate reliability indices
2. Do evaluate reliability with indices that are appropriate to the research question

Experimental cognitive researchers rarely confront issues of reliability in any comprehensive manner (e.g., if the scores on a measure

https://doi.org/10.1016/j.intell.2020.101440
Received 7 October 2019; Received in revised form 17 January 2020; Accepted 27 January 2020

Table 1
A list of Don'ts and Dos for laboratory studies examining intelligence.

Don't…	Do…
1. Don't disregard inadequate reliability indexes.	2. Do evaluate reliability with indices that are appropriate to the underlying construct.
4. Don't overgeneralize to external criteria without validation.	3. Do consider the validity of selected intelligence tests.
6. Don't use intelligence tests in a non-standard manner without norms and validation.	5. Do understand data are theory-laden.
7. Don't use extreme-group designs.	9. Do adjust correlations to account for restriction-of-range in talent.
8. Don't use difference scores.	12. Do pre-specify expected correlations.
10. Don't use samples too small to detect differences in correlations.	14. Do take account of speed vs. power of reference tests.
11. Don't confuse statistically significant with meaningful magnitude correlations.	15. Do take a faceted approach to assessing particular abilities.
13. Don't ignore common content as method variance.	17. Do take account of 'historical' and 'current' aspects of crystallized Intelligence.
16. Don't use 'indifference of the indicator' as a guide to test selection.	18. Do take account of underlying heterogeneity of the general intelligence construct.
20. Don't assume that all 'processing speed' ability assessments measure the same underlying factor.	19. Do use multiple tests to assess underlying ability factors.
21. Don't assume that adequate assessment with multiple measures requires excessive administration time.	

have too much variability for the comparisons of interest, the choice is often made to add additional task trials or more study participants to yield more stable [reliable] results). For example, in a review of experimental research, Green et al. (2016) found that only 5.7% of the recent articles in one journal and 11.7% in another reported reliabilities of their measures. Of these, most reports contained only internal consistency reliabilities, and only one study included an estimate of test-retest reliability. As Green et al. noted, "The results are consistent with our hypothesis that reliabilities are infrequently presented for experimental task scores." (p. 750).

When reliabilities are reported in this area of research, they are often startlingly low. For example, Hedge, Powell, and Sumner (2018) examined the reliability of a set of standard experimental tasks and noted that for "Eriksen Flanker, Stroop, stop signal, go/no-go, Posner cueing, Navon, and Spatial Numerical Association of Response Code....Reliabilities ranged from 0 to .82, being surprisingly low for most tasks given their common use." (p. 1166). Furthermore, reliabilities are frequently misinterpreted. For example, in another recent study, Craik, Bialystock, Gillingham, and Stuss (2018) reported a test-retest reliability of their alpha span test as "$r(42) = .61$ ($p < .001$). This highly significant correlation provides good evidence that the alpha span test is a reliable measure of working memory, at least in older adults." (pp. 145–146). However, the criterion for reliability is *not* a test against a null hypothesis of a zero correlation between a test administered on two occasions, as these authors imply. Rather, it is an estimate of the consistency of rank-ordering of individuals over some period of time. Ideally, if there are no changes in the underlying construct, their rank-ordering will not change, except for measurement error. So, the appropriate "standard" would be a correlation of $r = 1.0$.

Because test reliability sets an upper bound for validity (e.g., see Anastasi & Urbina, 1997), the choice of an appropriate indicator of test reliability is a critical component of determining the potential usefulness of the measure for the intended purpose. For example, internal consistency reliability for intelligence tests is often useful, at best, as an indicator of the homogeneity/heterogeneity of the underlying measure and can be too low or even too high, depending on the context in which the investigator is attempting to understand the construct under investigation. Coefficient α (also known as Cronbach's α) is the most frequently employed index of reliability, but it is very often misunderstood, and as a result, interpreted inappropriately. Coefficient α reflects the average of all possible split-half correlations among test items (Cronbach, 1951). As such, it is a statistic that essentially confounds the internal consistency reliability of test items with the homogeneity of the test, but as others have pointed out, not in a particularly straightforward manner, especially when there is more than one factor underlying the test items (e.g., see Schmitt, 1996; Sijtsma, 2009). In addition, α increases with the number of items in a test.

In practical terms, what this means for experimental research related to intelligence is that there is no specific, universal threshold for what constitutes a *good* or even an *acceptable* α value. Tests with narrow content or redundant items will have typically higher α indices, compared with tests of broad content. Thus, an evaluation of internal consistency reliability for a particular experiment and measure requires that the investigator determine a priori what is an appropriate value and evaluate the index against the expected value. Using an arbitrary threshold for all measures (e.g., $r_{xx'} = 0.70$), testing against a null hypothesis of zero reliability, or simply reporting a range of α values and proceeding to analyze and interpret the results *regardless* of the values of α (e.g., Miyake, Friedman, Emerson, Witzki, Howerter, and Wager, 2002; where reported reliabilities were as low as 0.31 and 0.42, p. 69), are all ill-advised, because each of these approaches will likely result in the inclusion and interpretation of measures with inadequate theoretical or measurement characteristics. When there is a mismatch between the expected measure reliability/homogeneity and the actual reliability/homogeneity index, there are likely both conceptual problems indicated and/or measurement difficulties (e.g., a measure with inappropriate low internal consistency reliability may have exceedingly low diagnostic power).

In addition, for many intents and purposes, α does not address the fundamental questions of reliability that need to be evaluated for measures of ability. R. L. Thorndike's (1947) depiction of the sources of variance in test performance provides an essential framework for understanding what type of reliability assessment is appropriate for a given measure (i.e., the dimensions of temporary vs. lasting and specific vs. general sources of variation). For example, α is not an appropriate reliability estimate for general intelligence, because in addition to the psychometric limitations discussed, it lumps all four quadrants of Thorndike's framework into 'true' score variance. Rather, the framework dictates that a suitable index of reliability for a measure of general intelligence is obtained by examining delayed test-retest alternate-form reliability (which puts temporary and specific, temporary and general, and lasting and specific variance into the error term, leaving only

lasting and general variance as 'true' scores, yielding a much more meaningful index of the test reliability, in contrast to an index of internal consistency (e.g., see Cronbach, 1960; [p. 138] illustration with the DAT Mechanical Reasoning Test, where 7% of test variance was found to be "temporary-specific", 8% "lasting-specific", 20% "temporary-general", and 65% of the variance was reliable ["lasting-general"] variance). To sum up, if reported at all, reliability statistics from laboratory studies seeking to link experimental cognitive constructs and intellectual ability are often misinterpreted, inappropriate, or both.

3. Validity

3. Do consider the validity of selected intelligence tests

In experimental cognitive psychology, there is ongoing debate about what particular laboratory tasks are suitable for assessing a particular construct (e.g., memory search speed, decision time, reaction time or movement time, verbal or spatial working memory). The "validity" of such tasks is established through discussion in the literature, leading to a set of relatively standardized procedures to establish that a particular set of task conditions yields an estimate of the speed or accuracy of responding underlying a particular construct. But in experimental studies considering intellectual abilities, the extant literature is largely devoid of explicit examination of content validity—that is, whether the content of the tests used to assess intelligence adequately sample the domain of intellectual ability. For example, a single figural reasoning test such as Raven's Progressive Matrices will often be used to measure general intelligence (e.g., Jaeggi, Buschkuehl, Jonides, and Perrig, 2008; see numerous other examples in the "brain training" literature, as reviewed by Simons et al., 2016).

3.1. Construct and criterion-related validity

Reports of construct validity are more frequent than reports of content validity, though few studies include any specification of criteria for acceptable convergent or discriminant validity, making such reports largely impressionistic at best. There is an occasional mention of "criterion validity," but indicators of criterion-related validity are usually not presented. For example, Redick, Broadway, Meier, et al. (2012) purported to address the "Construct and Criterion-Related Validity" (p. 167) of complex span tests. However, only correlations between working memory tests and Raven's Progressive Matrices and a Vocabulary test were reported. The values reported were correlations ranging from 0.20 to 0.53 for the Raven's, and from 0.09 to 0.15 for the Vocabulary test, with little critical evaluation to determine whether such values should be considered good or poor indicators of validity.

3.2. External (criterion) validity

4. Don't overgeneralize to external criteria without validation

Over the modern history of intelligence assessment, there have been two, sometimes partly divergent, streams of research, mainly identified with their initial proponents – Spearman (1904) and Binet (Binet & Simon, 1905). Spearman's approach, while initially focused on school performance (grades and ratings) as indicators of intelligence, eventually developed to mainly focus on articulating a theory of intelligence that was about finding fundamental properties of *g*, largely independent of external validation. The Binet approach was explicitly concerned with using assessments of intelligence for prediction of individual differences in academic success or failure. Followers of Spearman's approach have often focused on finding correlates of *g*, where *g* is operationalized by tests such as Raven's Progressive Matrices. But, there are two inherent limitations of this approach. First, scores on the Raven's test and similar non-verbal reasoning tests are imperfect measures of *g* (e.g., see Gignac, 2015), much in the way that any single test does not provide a robust assessment of the higher-order *g* factor. Second, in contrast to Binet's approach, use of tests like the Raven's is only moderately successful in predicting real-world criteria, such as academic or occupational success (e.g., see P.E. Vernon & Parry, 1949). For Binet and his followers, and especially for Wechsler, the construct of intelligence is *not* the score on an intelligence test, but rather is indicated by successful accomplishments in academic and occupational settings, and other real-world endeavors (e.g., see Wechsler, 1975).

Much experimental laboratory work tends to ignore external validation of measures (e.g., see Mook, 1983), which probably explains a certain affinity for approaches to intelligence assessment more in line with Spearman than with Binet. This isn't to say that such an approach is inherently flawed – it is an internally coherent approach within a reductionist framework. The potential problem is when a researcher attempts to generalize beyond the laboratory, on the basis of a highly limited assessment of intellectual ability, to make statements about the meaning of task/intelligence correlations for external criteria, such as academic or occupational success. Such statements mistake what is inherently a limited "predictor" variable for the "criterion" of intelligence assessment (e.g., for a discussion, see Ackerman, 2017). If the researcher desires external validation for experimental tasks to intellectual ability criteria, the researcher must actually conduct criterion-related validity research, by obtaining direct assessments of external criteria.

4. Data are theory-laden

5. Do understand that data are theory-laden

Thomas Kuhn (1977), physicist-turned-philosopher of science, proposed the theme of "paradigm" to describe the amalgamation of assumptions and methods used within individual laboratories. Although there was a lively argument in the literature about whether psychology could be considered paradigmatic, it is clear that within the domain of intelligence research, there are multiple paradigms, loosely associated with the Spearman and Cattell-Horn-Carroll (CHC) paradigms. The Spearman paradigm, as realized in modern experimental research, takes a reductionist view to intelligence, with its measurement typically operationalized by Raven's Progressive Matrices test or a similar figural reasoning test. The CHC paradigm is somewhat less unified, but it encompasses a consideration of a hierarchy consisting of a general intelligence factor, major group factors (e.g., fluid and crystallized intelligence), minor group factors, and so on (e.g., see McGrew, 1997). Variants of other hierarchical models (e.g., P. E. Vernon, 1950) provide for different identification of major and minor group factors, but most such variants consider these factors as themselves important indicators of intellectual ability, sometimes in the aggregate *more* important than the general intelligence factor.

The critical issue for research relating experimental research is that adoption of the Spearman approach, for example, implicitly disregards consideration of the possible relations between individual differences in task performance, and individual differences in major and minor group factors. With adequate sampling of the major and minor group factors one can estimate a general intelligence score, but one cannot examine the role of such factors when one only assesses a single measure that is thought to provide an estimate of general intelligence.

5. Standardized intelligence tests

6. Don't use intelligence tests in a non-standard manner without norms and validation

The traditional 'gold standard' assessments of intelligence (e.g., Stanford-Binet, Wechsler) are considered to be standardized tests, in that all of the examiner's words, apparatus, and procedure, and the scoring rules are fixed. The same is true for college admissions tests, such as the SAT or ACT, or tests administered in the manner dictated by test manuals. However, there are numerous variations that have been developed by various researchers that deviate from these procedures. Changing the context (e.g., administering the SAT in a low-stakes

laboratory environment), shortening the test, imposing different time limits, computerizing presentation of items, changing the instructions, and so on, may have unintended consequences for both the reliability and validity of the test. For example, several short-form versions of the Raven's Progressive Matrices have been developed (e.g., Arthur Jr., Tubre, Paul, and Sanchez-Ku, 1999; Biker et al., 2012), changing it from a two-part relatively untimed test to as few as 9 items with a strict time limit. Ceteris paribus (everything else being equal), having fewer test items will reduce the reliability of the test (e.g., by the Spearman-Brown prophecy formula). Arthur Jr. and Day (1994) reported a correlation of their short form and the standardized Raven's test to be $r = 0.66$, indicating that the two measures share less than 50% of their variance ($r^2 = 0.436$). Whether such a reduction of common variance (compared to test-retest reliability of the standardized Raven's test) is entirely attributable to the reduced reliability of the short form, or perhaps is at least partly attributable to the imposition of a speed requirement, is as yet unknown. But, if there is an increase in influence of speededness of processing, that could result in an artificially higher correlation between a short-form Raven's test and individual differences in performance on speeded experimental tasks that is *not* uniquely attributable to their respective common variance associated with *g*.

When tests are used in non-standardized fashion, two requirements need to be satisfied. First, norms need to be collected, so that there is an adequate basis for calculating correlations that are adjusted for restriction-of-range of talent (see below). Second, the test needs to be validated against other indicators for both convergent and discriminant validity. Convergent validity for such a test could be obtained from independent standardized measures of the ability under examination. Discriminant validity is especially important in this context, as it will allow for an assessment of 'extraneous' factors, such as processing speed or verbal comprehension (e.g., when abbreviated instructions are used).

6. Regression-to-the-mean

7. Don't use extreme-group designs

6.1. Extreme-group designs

This issue was first encountered in Wellman's (1940) study of the influence of a nursery school experience on intellectual abilities. In that study, Wellman reported observing increases in intellectual abilities on a low-IQ sub-sample, and a lack of increases in IQ in a high-IQ sub-sample. As noted by McNemar (1940) and Goodenough and Maurer (1940), the increases in IQ were entirely predictable because of the relatively low reliability of the IQ test for young children and the expected regression-to-the-mean from pre-test to post-test. Although most statistical experts recommend against using extreme-groups designs in general (e.g., see Preacher, Rucker, MacCallum, and Nicewander, 2005), at the very least, one must account for expected regression effects and interpret only those effects that significantly exceed the expected results on the basis of regression-to-the-mean effects.

There are additional concerns about valid interpretations of extreme-groups differences when the sample being investigated is already restricted in range-of-talent. For example, a so-called 'low-working memory' group sampled from a highly selective college or university is probably not all that representative of a 'low-working memory' group from an unselected sample. This was once a common practice in research on working memory (see numerous papers by Engle and colleagues before the mid-2000s, e.g., Kane, Bleckley, Conway, & Engle, 2001), and is still occasionally used in this area of research. For example, Hourihan and Benjamin (2010) classified University of Illinois undergraduate students as low-span or high-span based on their scores on the operation span task, even though the *median* SAT scores of this population are about the 90th percentile for the SAT (University of Illinois, 2020). In addition, by examining only extreme groups, one has greater difficulty in estimating the true score correlations between experimental variables and ability variables, unless statistical corrections are made. Under an extreme-groups design, any overall correlations observed across groups will typically over-estimate the underlying true score relations – correcting for the higher variance than the underlying population will result in an attenuated (closer to zero) estimated true score correlation.

6.2. Difference scores

8. Don't use difference scores

The statistical limitations of difference scores were first encountered in the study of learning abilities by Woodrow (1946). Woodrow calculated learning scores as the difference between initial performance on a set of tasks and final performance on the same tasks, after task practice. On finding extremely low correlations between these "learning" scores and extant ability measures (or even each other), Woodrow concluded that intelligence was not related to learning, and that there was no general or group learning factors. The problem with this assertion was that difference scores are notoriously unreliable (Cronbach and Furby, 1970). Even in the absence of any learning or treatment effect, the reliability of difference scores is determined by the reliability of the component (pretest and posttest) scores, and the correlation between the two scores. As the correlation between the scores increases, the reliability of the difference scores decrease; and as the reliability of the component scores themselves increase, the reliability of the difference scores increases, but ceteris paribus, increasing the reliability of the individual component scores will increase the correlation between the scores (Ghiselli, Campbell, & Zedeck, 1981). For example, if the reliabilities of the two component scores are $r_{xx'} = 0.8$ and $r_{yy'} = 0.8$, and the correlation between the two component scores is r_{xy} 0.70, the reliability of the difference score is $r_{dd'} = 0.33$, a value that would render such a score without merit for further correlations with other variables. The use of measures based on difference scores has been common in laboratory studies of intelligence, and not surprisingly, reliability estimates are often quite low for these measures (for a recent example, see Kalra, Gabrieli, and Finn, 2019).

7. Restriction-of-range in talent

9. Do adjust correlations to account for restriction-of-range in talent

When a researcher chooses a sample that is restricted in range-of-talent – a typical phenomenon when the study participants are students from a moderate to highly selective college/university, correlations are expected to be attenuated than would be observed in the population-at-large. Corrections can be made to the correlations, on the basis of comparing the variances of the sample on the variables of interest to the variances in the population at-large (e.g., see Ghiselli et al., 1981). Of course, if there are no representative data from the population at large on the variables, it is not possible to provide an accurate estimate of true-score correlations. As an example, as yet, there are no norms based on representative data for widely used working memory tasks such as operation span, running span, and visual arrays; the vast majority of data for these tasks are from undergraduate students who were selected for college admission based on intellectual talent (i.e., SAT or ACT score).

A related issue concerns the restriction-of-range that is the result of explicit selection, such as occurs when intellectual abilities or measures that are substantially related to intelligence (e.g., SAT, ACT, grade point average) are used to determine which individuals are in the sample to begin with. Fortunately, although such corrections are rarely, if ever, used in laboratory experimental studies, formulas also exist for estimating true score correlations under explicit selection conditions (e.g., see Boone & Lewis, 1978; Gulliksen, 1950; Thorndike, 1949; Wiberg & Sunström, 2009). Failure to attend to these issues may result in a substantial underestimate of the relations between experimental tasks and measures of intellectual abilities. But, more seriously, if particular

intellectual abilities are more highly associated with the selection variables than other abilities, failure to account for explicit selection may result in differences in correlations between the experimental task and different ability measures that are statistical artifacts of the explicit selection process.

8. Differences between correlations

10. Don't use samples too small to detect differences in correlations

8.1. Detecting differences in correlations

Often, an investigator will wish to show that there is a difference between sets of correlations between experimental task performance and one or more ability measures. Detterman (1989) provided an incisive analysis of the difficulty in obtaining statistically significant results with such an analysis. The test for differences between correlations has a much lower statistical power than testing for significance of individual correlations against a null hypothesis ($r = 0$) or specific value. To illustrate, assume that an investigator finds that general intelligence correlates $r = 0.5$ with one experimental task and $r = 0.3$ with another experimental task, and wishes to test the difference between the correlations. According to Detterman, it would require a sample size of at least 200 participants to detect a difference between correlations of this magnitude, far more participants than are sampled in a typical laboratory experimental study. For smaller correlations, the sample sized required increases exponentially (e.g., 1000 participants are needed to detect a difference of 0.1 when the smallest correlation is $r = 0.3$).

It is also critical to note that when one correlation is significantly different from zero whereas another correlation is not, this doesn't mean that there is a significant *difference* between the two correlations. In fact, it is not at all uncommon for the difference between two such correlations to be non-significant. As one example, in a review, Ericsson (2014) cited the finding by Ruthsatz, Detterman, Griscom, and Cirullo (2008) that scores on a test of reasoning ability (Raven's Progressive Matrices) correlated significantly with a measure of musical achievement in novice musicians ($r = 0.25$) but not in samples of highly skilled musicians ($rs = 0.12$ and 0.24) as support for his hypothesis that the acquisition of skill attenuates the effect of cognitive ability on domain-specific performance. However, this conclusion is not warranted, because the correlations are not significantly different from each other across skill level.

9. Significant vs. meaningful correlations

11. Don't confuse statistically significant with meaningful magnitude correlations

12. Do pre-specify expected correlations

Over the past few decades, scholarly organizations and journal editors have encouraged researchers to move beyond null hypothesis significance testing, in favor of reporting effect sizes (e.g., Wilkinson and Task Force on Statistical Inference, 1999). This movement is especially important for laboratory studies of experimental tasks and intellectual abilities, because it may force researchers to consider what the magnitude of associations between such variables actually means. The researcher needs to ask: "Is a correlation of $r = .40$ between such variables meaningful, or is it an over- or under-estimate of what should be expected, on the basis of existing theory?" Too often, researchers refer to similar correlations as (for example) "large," "medium," or "substantial," without any reference to what was expected or what is consistent with prior research or theory. By pre-specifying an expected magnitude of correlations, it will be much more likely that the study results will be informative and less subject to ad hoc explanations. Moreover, such an analysis might spur the researcher to additionally report estimated true-score correlations (based on corrections for restriction of range in talent, reliability of the measurements, and so on), in addition to raw observed correlations.

A corollary to recommendation #11 is "don't ignore correlations that are too small to be meaningful and moreover are not even statistically significantly different from zero," when adequate sample sizes are obtained. For example, Fournier-Vicente, Larigauderie, and Gaonac'h (2008) administered a battery of "19 tasks thought to assess ... six executive functions" (p. 35) to a sample of $N = 180$ undergraduate students. A model was constructed to represent the six executive function latent factors of executive function. Although the authors reported that the model did not fit the data, changes to the model specifications yielded a "a five factor model [that] fits the data well" (p. 42). The difficulty is that even though the structural equation model (SEM) could not be rejected on the basis of specified criteria, the input matrix illuminates that, in Horn's terms, the factors were essentially 'slicing smoke' (see Horn, 1989), particularly when one takes account of the underlying correlations. We calculated the mean and median values of the 19 variable correlation matrix and found that the mean correlation was $r = 0.11$, and the median correlation was $r = 0.09$. What's more, 22% of the correlations were actually negative, which should give one pause, given the ubiquitous positive manifold found across the ability domain. A calculation of squared multiple correlations (SMCs) for each variable was derived to estimate the communality for each variable in the set of tasks. Thirteen of the 19 variables had SMCs below 0.04, meaning that less than 4% of the variance of these variables was common to the matrix, and $>$ 96% of the variance was either unique to each task or was error. Although it is always useful to remember that factor solutions are not unique (i.e., there is an infinite number of other models that equally well fit the data), when a correlation matrix that is largely made up of essentially miniscule magnitude correlations, there is even more potential for misleading conclusions from such analyses. We suggest that it is doubtful that one could derive a meaningful and valid factor-analytic model inference from variables with virtually no common variance.

When it comes to SEM models in general, variables with low correlations (because of low reliabilities [see "Don't" #1 above] or measures with otherwise low levels of *SD* (inter-individual differences) result in both unpredictable effects on the relations of predictors and criterion variables (e.g., see Maruyama, 1998) or low power to reject models because of illusory indicators of good data-model fit (e.g. see Hancock & Mueller, 2011), respectively.

10. 'Method' factors

As noted by Campbell and Fiske (1959), "Each test or task employed for measurement purposes is a *trait-method unit*, a union of a particular trait content with measurement procedures not specific to that content." (p. 81). In the assessment of intellectual abilities, "method" can refer to a variety of different characteristics of a particular test, such as the kinds of question formats (multiple choice, open-ended, sentence completion), administration procedures (e.g., speed vs. power tests), item stimuli (e.g., words, numbers, familiar pictures vs. novel polygons), computerized vs. paper and pencil vs. one-on-one oral examination, and so on. Different tests of reasoning abilities, for example, can be found with any of the above different characteristics. The general expectation, however, is that when two tests share any underlying method characteristic, scores on the tests may correlate more highly than would be expected based solely on their common variance on the underlying trait constructs; that is, they share variance on both trait factors and method factors. When experimental tasks share any of these method characteristics with the accompanying intellectual ability measures, they may have spuriously higher correlations with one another. Failure to take account of common method variance may yield inappropriate conclusions about the common task-trait variance. For example, in research on working memory (e.g., Kane et al., 2004), it has been a common practice to use "complex span" tasks having similar procedures to measure working memory capacity (i.e., interleaved

storage and processing tasks). Consequently, a latent factor comprising these measures will reflect not only construct-related variance, but also method variance, and it will not be possible to determine whether correlations of the latent factor with other factors are at least in part due to this method variance.

11. Content as overlapping 'Method' factor

13. Don't ignore common content as method variance

One of the reasonably well-replicated findings from research conducted under the heading of "elementary information process" correlates of intelligence (e.g., Ackerman, 1986; Kyllonen, 1985), is that common stimulus content (e.g., numbers as stimuli/test items) on both the task and ability assessments will lead to higher correlations than when there are different stimulus contents (e.g., a spatial stimulus task and a verbal ability test). This phenomenon could be considered an example of common method variance, or it could be considered as representing something closer to overlapping stimulus bonds (e.g., Thompson, 1919; see also Kovacs & Conway, 2016). The main concern regarding this phenomenon is to make sure that whatever common variance is found between task and ability measure takes into account the fact that the source of communality may not be entirely attributable to the underlying task or ability factor(s), if those entities also encompass other stimulus content types. For example, if the investigator only examines a working memory task with spatial content and a nonverbal (spatial stimulus) intelligence measure, but attempts to describe the resulting correlation as revealing the common variance between "working memory" and "intelligence," there is likely to be an overestimate, compared to situations where *different* stimulus contents are used for the same kind of evaluation.

12. Speed vs. power method issues

14. Do take account of speed vs. power of reference tests

As the information processing framework in experimental psychology grew out of information theory (e.g., see Atkinson & Shiffrin, 1967; Attneave, 1959; Shannon and Weaver, 1949), the two major dependent variables for such investigations were the speed and accuracy of responding, often with a major emphasis on speed.[1] Assessments of intellectual ability, in contrast, often take a different approach to determining level of performance. There are a number of straightforward speed tests, such as those in the domains of perceptual speed and psychomotor ability assessments, where items are typically low in difficulty, and the number of completed items in a fixed time period represent the estimated ability level of the examinee. But, the majority of intelligence assessments either are exclusively power tests (with easy items at the beginning of the test, and items of increasing difficulty as the test proceeds), with no time limits or very generous time limits, or some combination of speed and power (items of increasing difficulty, but strict time limits). Theorists have indeed argued about whether intelligence as a construct should include the speededness of cognitive processing as a fundamental element (e.g., Sternberg, 1986, Vernon, Nador, and Kantor, 1985).

Regardless of the philosophical disagreements about the underlying nature of the construct of intelligence, when speed (restrictive time limits) is used in intelligence assessments, along with speed as the key dependent variable in the experimental tasks, an increase in communality across the tasks will exist, in comparison to a speeded experimental task and a true power-test format for intellectual ability assessment. This issue is essentially the same as other common method concerns – inferences about the commonality of the underlying experimental construct and the ability construct may result in overestimates, mainly as a result of these common methods.

13. Approaches to assessing intellectual abilities

The following is a set of recommendations, and their underlying justifications, for assessing intellectual abilities in the context of laboratory experimental studies aimed at determining the relations between tasks and intellectual abilities. Examples are provided to underline these recommendations.

13.1. 'Factor-pure tests' versus a faceted approach

15. Do take a faceted approach to assessing particular abilities

One of the mismatches between experimental psychology approaches to psychological constructs and differential approaches to the assessment of intellectual abilities lies in the underlying phenomena of interest. Experimental researchers can arguably determine fundamental constructs, such as the amount of time required to search memorized words or a display of various symbols, by calculating differences between task conditions, averaged across task trials, participants, or both (e.g., Hick, 1952; Hyman, 1953; Sperling, 1960). The parallel approach to determining individual differences in a single underlying ability would be to find a test that is "factor pure," that is, where the test completely identifies the underlying ability construct. In an experiment where experimental task variables are correlated with individual scores on a single homogeneous test of intellectual ability (e.g., Raven's Progressive Matrices), there is either an implicit or explicit assumption that the test is a "factor pure" assessment of the construct (e.g., Gf or Spearman's *g*).

In order for a factor pure test to actually exist, scores on the test would necessarily load only on a single underlying ability factor, within a wide sampling of administered ability tests. Unfortunately, empirical evidence and intelligence theory both indicate that this is not a likely outcome of such an analysis. The hierarchical structure of intellectual abilities that represents the dominant CHC framework for intelligence indicates that the higher-order factors are a function of common variance among minor group factors and major group factors. This means that each test has some valid component of individual differences variance in these lower-order factors. Even Raven's Progressive Matrices, which is often claimed to be highly representative of Spearman's *g* or Gf, shares substantial common variance with lower-order abilities of reasoning and factors of spatial ability (e.g., see Burke, 1958; Gignac, 2015). The hierarchical representation of abilities involves a partitioning of the reliable variance of each test into specific variance (that is, reliable variance that is unrelated to any other test), common variance with lower order factors, and common variance with higher-order factors.

Therefore, if one wishes to compare individual differences in experimental task performance with estimates of a particular intellectual ability, one must provide a more robust accounting of that intellectual ability, that is less influenced by variance that is specific to the test itself, and if desired, is less influenced by lower-order factor variance. The solution to this problem was suggested by Humphreys (1962), with an orientation that was suggested by Guttman's (1954-1955) facet theory. Because, by definition, specific variance on each test is uncorrelated with specific variance on all other tests, the optimal approach to estimating individual differences on a particular underlying ability is to *maximize* heterogeneity among tests that purport to measure the same ability. More recently, this idea has been discussed in terms of the concept of "controlled heterogeneity" (see Little, Lindenberger, and Nesselroade, 1999).

Consider the example of an assessment of reasoning ability. The assessment should sample across item content (e.g., "numbers, words, figures, and photographs") and item format (e.g., "analogies, series, and

[1] Because speed and accuracy are considered to have a non-linear relationship, study participants are often instructed to keep a constant high level of accuracy, and to respond as quickly as possible without sacrificing accuracy, see Wickelgren (1977).

classification"; Humphreys, 1962, p. 482). By extension, an optimal estimate of general intelligence would sample across all different test contents and item formats that underlie the construct. Fundamentally, aggregating across tests with each type of content and format will result in the specific sources of variance in each test cancelling out, leaving one with a robust assessment of the underlying ability construct. The faceted approach also increases the breadth of the assessment, a factor that will likely lead to increased validity for task-ability correlations (see Wittmann & Süß, 1999 for the concept of "Brunswik Symmetry"). Studies of cognitive aging literature by Salthouse and colleagues illustrate this approach to assessing intellectual abilities. In these studies, participants complete a "reference battery" consisting of multiple tests to measure each of several established cognitive constructs, such as reasoning ability, memory, processing speed, and vocabulary (e.g., Salthouse, Pink, and Tucker-Drob, 2008; Siedlecki & Salthouse, 2014). The tests vary in both item content and format, permitting greater confidence that latent variables reflect the hypothesized cognitive constructs of interest rather than facility in performing specific types of tests.

13.2. Fluid intelligence and the so-called "indifference of the indicator"

16. Don't use 'indifference of the indicator' as a guide to test selection

A fall-back position, whether implicit or explicit, for many experimental studies, is that it doesn't much matter which test or tests are used to assess intelligence or fluid intelligence (Gf) or *g*, because of Spearman's "theory" of the "indifference of the indicator." Briefly, Spearman claimed that "for the purpose of indicating the amount of *g* possessed by a person, any test will do just as well as any other, provided only that its correlation with *g* is equally high" (Spearman, 1927, p. 197). Although there are some differential psychologists who agree with this claim, there have been substantial objections, best articulated by Horn and McArdle (2007) on both philosophical grounds ("it assumes what it is trying to prove", p. 222), and on empirical grounds (by showing clear discrepancies in the constitution of *g* from different data sets and even the same variables measured on different samples). As noted by Horn and McArdle, a general factor derived from a sample of spatial content tests will be fundamentally different from a general factor derived from a sample of verbal content tests. Or, in more specific terms, the *g* that is highly determined by the Ebbinghaus Completion Test (Ebbinghaus, 1896–1897; a test of verbal memory and fluency; see Ackerman, Beier, and Bowen, 2000) is inherently different in content and both construct and criterion-related validity from a *g* that is highly determined by the Raven's Progressive Matrices Test.[2]

13.3. Crystallized intelligence

17. Do take account of the "historical" and "current" aspects of crystallized intelligence

Assessment of crystallized intelligence (Gc) is complicated by three different aspects. The first is that, as with other higher-order ability factors, Gc is made up of diverse underlying lower-order ability factors. Sampling only one aspect of Gc (such as verbal comprehension or general information) results in a compromised assessment similar to measuring only a single aspect of Gf. A second complication comes from the developmental aspect of Gc, as noted by Cattell (1971-1987). That is the distinction between "current" Gc and "historical" Gc. In Cattell's view, historical Gc is represented by content that was learned in childhood or adolescence, while current Gc is represented by more recently learned knowledge and skills. Assessments that often are presented as measures of Gc, such as the SAT verbal composite (which is made up of reading, writing and language content) represents material that may have been acquired years prior to the administration of the test, and moreover, may not consider Gc content that falls outside of the common core curriculum of a middle school or high school curriculum, such as knowledge of different languages, literature, and so on. The third aspect of Gc that complicates assessment is related to the lack of consideration of "current" Gc. That issue is the diversity of knowledge and skills that are theorized to represent the construct. For adults, that means that substantial components of Gc include domain knowledge that is not common to individuals with different educational, occupational or avocational experiences and expertise. Ultimately, assessments of Gc that concentrate on the "historical" aspect of the construct need to recognize that such an assessment substantially under-samples the overall Gc construct. One alternative is to conduct a wider sampling of "current" Gc, such as might be achieved with administration of a wide variety of domain knowledge scales (e.g., see Ackerman, 2000), or to include more specialized educational knowledge and occupational/avocational knowledge assessments.

13.4. Measuring general intelligence

18. Do take account of underlying heterogeneity of the general intelligence construct

As noted earlier, there are multiple major theoretical frameworks for the hierarchical structure of intelligence, such as the CHC framework. It is useful to note that for laboratory assessments of intellectual abilities, where correlations between experimental tasks and intellectual abilities are concerned, the overall goal of a particular study is *not* typically to provide an overall assessment of all of the major intellectual abilities demarcated by an overarching theory, but rather to make comparisons with particular abilities, whether higher or lower on the hierarchy.

If one is interested in correlating task performance or related variables to general intelligence, the best approach is to administer a battery of tests of tests that are heterogeneous in content across all the major group or higher-order factors (e.g., as represented in the CHC theory; fluid intelligence, crystallized intelligence, general visual intelligence, quantitative ability, speed, short and long-term memory, and so on). Omnibus intelligence tests, such as the Stanford-Binet or Wechsler tests, are generally considered to be the gold standard for general intelligence assessments, but these require extensive administration time and individual testing. There are group tests that attempt to provide similar sampling, such as the Wonderlic (1981), but it is important to recognize that any such assessment will make compromises in terms of the heterogeneity of content, reliance on speeded tests, and potentially a more fundamental concern, require that the individual know how to read (and sometimes know how to write). As noted by Carroll (1982), these group tests make an unstated assumption that individual differences in the reading and writing skills necessary to perform the test are unrelated to individual differences on the construct to be measured – an assumption that may not be viable, depending on the nature of the participant sample.

13.5. Measuring content abilities directly

19. Do use multiple tests to assess underlying content ability factors

With respect to content abilities (i.e., verbal, numerical, spatial), the CHC framework provides only separate identification of numerical (Gq, quantitative reasoning/knowledge) and spatial (Gv, visual intelligence/processing). As with other hierarchical theories of intelligence (e.g., P. E. Vernon, 1950), each of the general abilities can be subdivided into constituent narrower abilities (e.g., see Lohman, 1979 for an explicit breakdown of spatial abilities). To achieve a robust estimate of one of these higher-order abilities, it is necessary to administer multiple ability

[2] It may be interesting to note that Krueger and Spearman (1907) reported that the Ebbinghaus Completion Test had a loading of 0.97 with *g* – a value higher than any other measure reported to that date or afterwards by Spearman and his colleagues.

assessments that are heterogeneous on the lower-order factors that make-up the higher order factor. Otherwise, over-sampling or under-sampling the constituent factors runs the risk of not fully identifying the higher-order factor, and only estimating one or more of the lower-order factors. Simply naming a test as representative of a higher-order factor and using the higher-order factor name as the identified individual-differences construct is sloppy at best, and misleading at worst. An example of this practice comes from a study of brain training by Jaeggi and colleagues (Jaeggi et al., 2008), who equated scores on a single test of figural reasoning with Gf.

Verbal content in the CHC framework is encompassed in Gc, which also includes vocabulary and production fluency, but also in a general reading and writing ability (*grw*). An assessment of a broad verbal ability factor is possible, but an assessment of *grw* would not ordinarily include fluency or knowledge measures. Thus, an adequate assessment of verbal content ability would probably need to sample across the verbal content of Gc ability *and* assessments of reading and writing abilities.

As an example, an attempt to measure the three major content abilities might include three or four measures of each, which are heterogeneous on the lower-order factors. An assessment of math ability might include a measure of math reasoning (e.g. number series), math knowledge (e.g., math word problems), and geometry. Such an assessment would have the additional advantage of utilizing three different stimulus types (numbers, words, figures), which means that an aggregated composite of math ability would be expected to be less influenced by method variance. Assessments of spatial ability would be less likely to be able to be entirely heterogeneous with respect to content, but one could construct a battery of tests that included spatial reasoning (e.g., spatial analogies), speeded rotation (e.g., flags), spatial visualization (e.g., the Vandenberg & Kuse, 1978), and spatial word problems (e.g., the Verbal Test of Spatial Ability; see Ackerman & Kanfer, 1993). Assessments of Verbal ability also have limitations on stimulus types, but an adequate battery of measures that is relatively heterogeneous on content and processing components might include a verbal analogies (reasoning), vocabulary (verbal knowledge) and a completion test similar to that developed by Ebbinghaus.

One positive example of using multiple tests to establish underling intellectual ability factors in the context of examining correlations with elementary processes is by Van Dyke, Johns, and Kukona (2014). Although the study had a relatively modest sample size ($N = 65$), the authors sampled from the community, rather than from pre-selected college/university students. They administered 24 ability measures to assess “print mapping, reading skill, oral language use, memory” and they included a norm-reference intelligence measure (the Wechsler Abbreviated Scale of Intelligence), along with measures of working memory and memory for serial order. As a result, it was first possible for the authors to determine the representativeness of the sample, with respect to mean and variance of IQ (Mean = 95.88, sd = 15.08). Extensive analyses were then conducted to determine that working memory capacity was a “spurious determinant of poor comprehension” (p. 384). Such results probably would have been less likely to be obtained if the battery of reference ability tests had not been so extensive.

13.6. Measuring speed abilities

20. Don't assume that all ‘processing speed’ ability assessments measure the same underlying factor

Given the inherent speededness of many experimental tasks, an investigator might desire to directly estimate individual differences in the speed components of intellectual abilities as a separate variable, in order to determine the level of overlap between underlying general intelligence and an experimental task, with speed ability partialled out. Numerous investigators have made this kind of comparison, using one or more measures of “processing speed” as the indicators of speeded intellectual abilities. Although a “general speededness” ability was posited by Horn (e.g., Horn, 1965), more recent research has established that there are perhaps four related perceptual speed (PS) factors (e.g., Ackerman, Beier, & Boyle, 2002): PS-Scanning, PS-Pattern Recognition; PS-Memory and PS-Complex. A single test or even two tests of PS abilities is unlikely to well-identify a higher-order PS ability. Moreover, the degree of overlap between experimental task variables and PS ability estimates will likely partly depend on the overlap of underlying common content (stimulus type) and processes for the respective measures. A robust estimate of a general PS factor would require assessments of tests that represented at least the three main PS factors of Scanning, Pattern Recognition and Memory.

If one is interested in separating the accuracy from speed components of intelligence, perhaps the most promising approach would be to either administer tests with differing speed requirements (ranging from pure power tests to moderately to highly speeded tests of the same content – see Estrada, Román, Abad, & Colom, 2017; Lord, 1956), in order to de-couple the speed from level components, or statistically partial out estimates of perceptual speed and psychomotor abilities from correlations between content/general abilities and the target experimental tasks, or partial out content/general abilities from correlations between PS and psychomotor abilities and the experimental tasks (e.g., see Ackerman et al., 2002).

14. Contrasting a single ability measure with multiple ability measures

21. Don't assume that adequate assessment with multiple measures requires excessive administration time

Ackerman et al. (2000) administered a large set of intellectual ability tests, including the Raven's Advanced Progressive Matrices (APM) test, and a battery of 7 working memory (WM) ability tests to a group of 135 young adults, in order to determine the relationships between intelligence, perceptual speed, and working memory ability. The APM was administered in the manner specified by the manual (Raven, Court, and Raven, 1977), where there are oral instructions lasting 4.5 min, and a test of two parts – the first part has 12-items with a 5-min time limit, and the second part has 36 items, with an extended time limit of 40 items, for a total of 49.5 min administration time. Individual test part scores for the Raven's APM and a unit-weighted z-score composite of WM tests were correlated $r = 0.334$ for Part 1, $r = 0.470$ for Part 2, and $r = 0.475$ for the total score (Part 1 + Part 2). The Raven's total score correlated $r = 0.330$ with a Verbal Ability composite, $r = 0.352$ with a Math Ability composite, and $r = 0.694$ with a Spatial Ability composite, consistent with the extant literature that has described the test as much more highly related to spatial ability than other abilities.

In contrast, selecting a representative set of tests across the Verbal, Math, and Spatial domains could result in a more representative and robust estimate of general intelligence, and equivalent correlations with the WM composite, with a significant savings in administration time. Selecting one test from each domain with the shortest administration time to provide a general ability composite would include ETS Word Beginnings (7.3 min total administration time), Thurstone's Number Series Test (6.2 min) and Lohman's test of Spatial Orientation (see Ackerman & Kanfer, 1993) (6.5 min), for a total of a 20 min administration time. A unit weighted z-score composite of these three tests correlated $r = 0.437$ with the WM composite (nearly identical to the Raven's total score correlation of r 0.475), while this general ability composite notably only correlated $r = 0.455$ with the Raven's total score!

The study included 7 Verbal ability tests, 7 Math ability tests and 5 Spatial ability tests. Selecting one test each that loaded most highly on respective factors of these content abilities (Completion Test for Verbal, Arithmetic for Math, and Paper Folding for Spatial) would result in a battery of tests that requires 33 min to administer. A general ability composite of these three tests yielded a correlation with the WM ability

composite of $r = 0.567$, yet the correlation between this general ability composite and the Raven's total score was $r = 0.606$.

Tests that represent three of the four underling factors of Perceptual Speed (PS) ability (Memory, Scanning, Pattern Recognition) also account for variance in the WM ability composite, with a minimal additional investment of administration time, as these tests are very brief. Administrating one test from each factor (Digit/Symbol, Name Comparison, and Summing to 10) would take less than 20 min. Yet a composite of these tests correlated $r = 0.384$ with the WM ability composite, and correlated non-significantly with the Raven's total score ($r = 0.096$). A total ability composite of the shortest content ability tests and the three perceptual speed ability tests correlated $r = 0.514$ with the WM ability composite, and $r = 0.321$ with the Raven's total score. This battery of tests would only take about 40 min to administer, 9.5 min less than administering the Raven's APM.

Similar results were obtained in a study of $N = 117$ young adults by Ackerman and Beier 2007; (Study 3). From that study, a general ability composite of three brief tests representing Verbal (Vocabulary), Numerical (Number Series) and Spatial (Spatial Orientation) abilities with a total administration time of 20 min yielded a correlation of $r = 0.639$ with a composite of six Working Memory ability tests, while the Raven's APM total test score correlated $r = 0.560$ with the Working Memory composite. Including three of the same PS tests as in the earlier study with the content ability tests in a single composite yielded a correlation with the WM ability composite of $r = 0.762$, while correlating $r = 0.524$ with the Raven's APM total test score.

Such results indicate that a more robust measure of general intellectual ability than the Raven can be obtained with a sampling of content ability tests, with the additional advantage that with a shorter administration time for these tests, it is possible to more widely sample other abilities (such as PS abilities) that would allow for an expanded assessment of construct validity for other abilities (e.g., Working Memory).

15. Conclusions

Many of the “don'ts” and too few of the “do's” can be found in the extant literature on the relationship between constructs from experimental cognitive psychology and intellectual ability. Although they are prominent in laboratory experimental studies that attempt to relate task performance to individual differences in intellectual abilities, they are pervasive in the wider literature beyond the study of intelligence. It is important to note that there are few, if any, “perfect” studies in the literature, because of limitations in funding, time, effort, availability of optimal samples, and so on. The goal of this paper was to provide best-practice advice, in the hope that new studies will provide more definitive research results in the continuing efforts to integrate experimental and correlational approaches to understanding human intelligence. Thankfully, many of the do's and don'ts we have listed in this paper, such as correcting for restriction-of-range in talent, or using a small set of content measures to generate a robust general ability composite, require few, if any, additional resources, over and above current and past practices. Some “do” recommendations indeed require additional efforts, such as obtaining norms and validity estimates for non-standard test administrations. However, these kinds of investments are expected to provide long-term payoff in heuristic terms, by providing improved reference indicators for future investigations.

References

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

Ackerman, P. L. (2000). Domain-specific knowledge as the “dark matter” of adult intelligence: gf/gc, personality and interest correlates. *Journal of Gerontology: Psychological Sciences, 55B*(2), P69–P84.

Ackerman, P. L. (2017). Adult intelligence: The construct and the criterion problem. *Perspectives on Psychological Science, 12*(6), 987–998.

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*, 413–432.

Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2000). Explorations of crystallized intelligence: Completion tests, cloze tests and knowledge. *Learning and Individual Differences: A Multidisciplinary Journal in Education, 12*, 105–121.

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

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.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). New York: Prentice Hall.

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), 397–403.

Arthur, W., Jr., 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*, 354–361.

Atkinson, R. C., & Shiffrin, R. M. (1967). Human memory: A proposed system and its control processes. In K. W. Spence, & J. T. Spence (Eds.). *The psychology of learning, remembering, and forgetting: Proceedings of the second conference* (pp. 89–195). New York: New York Academy of Science.

Attneave, F. (1959). *Applications of information theory to psychology: A summary of basic concepts, methods, and results.* New York: Henry Holt.

Biker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven’s standard progressive matrices test. *Assessment, 19*(3), 354–369.

Binet, A., & Simon, T. (1905). New methods for the diagnosis of the intellectual level of subnormals. *L’Année Psychologique, 11*, 191-244. (Translated by Elizabeth S. Kite and reprinted in In J. J. Jenkins, & D. G. Paterson (Eds.). *Studies of individual differences: The search for intelligence* (pp. 90–96). NY: Appleton-Century-Crofts (1905)).

Boone, J. O., & Lewis, M. A. (1978). *The development of the ATC selection battery: A new procedure to make maximum use of available information when correcting correlations for restriction in range due to selection (FAA-AM-78-36).* Washington, DC: U.S. Department of Transportation, Federal Aviation Administration.

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Carroll, J. B. (1982). The measurement of intelligence. In R. J. Sternberg (Ed.). *Handbook of human intelligence* (pp. 29–120). Cambridge: Cambridge University Press.

Cattell, R. B. (1971-1987). *Abilities: Their structure, growth and action. [revised and reprinted as Intelligence: Its structure, growth, and action].* Amsterdam: North-Holland (1971-1987)).

Craik, F. I. M., Bialystock, E., Gillingham, S., & Stuss, D. T. (2018). Alpha span: A measure of working memory. *Canadian Journal of Experimental Psychology, 72*(3), 141–152.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671–684.

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). New York: Harper & Brothers.

Cronbach, L. J., & Furby, L. (1970). How we should measure “change” – or should we? *Psychological Bulletin, 74*, 68–80.

Detterman, D. K. (1989). The future of intelligence research. *Intelligence, 13*, 199–203.

Ebbinghaus, H. (1896–97). Über eine neue Methode zur Prüfung geistiger Fähigkeiten und ihre Anwendung bei Schulkindern. [On a new method for testing mental abilities and its use with school children.]. *Zeitschrift für Psychologie und Psysiologie der Sinnesorgane, 13*, 401–459 (Trans. by Wilhelm, 1999).

Ericsson, K. A. (2014). Why expert performance is special and cannot be extrapolated from studies of performance in the general population: A response to criticisms. *Intelligence, 45*, 81–103.

Estrada, E., Román, F. J., Abad, F. J., & Colom, R. (2017). Separating power and speed components of standardized intelligence measures. *Intelligence, 61*, 159–168.

Fournier-Vicente, S., Larigauderie, P., & Gaonac’h, D. (2008). More dissociations and interactions within executive functioning: A comprehensive latent-variable analysis. *Acta Psychologica, 129*, 32–48.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* New York: Freeman.

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.

Goodenough, F. L., & Maurer, K. M. (1940). The relative potency of the nursery school and the statistical lab in boosting the IQ. *Journal of Educational Psychology, 32*, 541–549.

Green, S. B., Yang, Y., Alt, M., Brinkley, S., Gray, S., Hogan, T., & Cowan, N. (2016). *Psychonomic Bulletin & Review, 23*, 750–763.

Gulliksen, H. (1950). *Theory of mental tests.* New York: John Wiley & Sons.

Guttman, L. (1954-1955). An outline of some new methodology for social research. *The Public Opinion Quarterly, 18*(4), 395–404.

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement, 71*(2), 306–324.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavioral Research, 50*, 1166–1186.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology, 4*, 11–26.
Horn, J. L. (1965). *Fluid and crystallized intelligence: A factor analytic study of the structure among primary mental abilities.* Ann Arbor, MI: University Microfilms International.
Horn, J. (1989). Models of intelligence. In R. L. Linn (Ed.). *Intelligence: Measurement, theory and public policy* (pp. 29–73). Urbana, IL: University of Illinois Press.
Horn, J. L., & McArdle, J. J. (2007). Understanding human intelligence since Spearman. In R. Cudeck, & R. C. MacCallum (Eds.). *Factor analysis at 100: Historical developments and future directions* (pp. 205–247). Mahwah, NJ: Erlbaum.
Hourihan, K. L., & Benjamin, A. S. (2010). Smaller is better (with sampling from the crowd within): Low memory-span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1068–1074.
Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist, 17*, 475–483.
Hunt, E., Frost, N., & Lunneborg, C. (1973). Individual differences in cognition: A new approach to intelligence. In G. Bower (Vol. Ed.), *Advances in learning and motivation. vol 7. Advances in learning and motivation* (pp. 87–122). New York: Academic Press.
Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology, 30*, 188–196.
Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences. 105. Proceedings of the National Academy of Sciences* (pp. 6829–6833).
Kalra, P. B., Gabrieli, J. D. E., & Finn, A. S. (2019). Evidence of stable individual differences in implicit learning. *Cognition, 190*, 199–211.
Kane, M. J., Bleckley, M. K., Conway, A. R., & Engle, R. W. (2001). A controlled-attention view of working memory capacity. *Journal of Experimental Psychology: General, 130*, 169–183.
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*, 189–217.
Kovacs, K., & Conway, A. R. A. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry, 27*, 151–177.
Krueger, F., & Spearman, C. (1907). Die korrelation zwischen verschiedenen geistigen leistungsfähigkeiten. *Zeitschrift fur Psychologie und Physiologie der sinnesorgane. 44. Zeitschrift fur Psychologie und Physiologie der sinnesorgane* (pp. 50–114). Trans. by W. W. Wittmann.
Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. *The essential tension: Selected studies in scientific tradition and change* (pp. 320–339). Chicago: University of Chicago Press.
Kyllonen, P. C. (1985). *Dimensions of information processing speed (AFHRL-TP-84-56).* Brooks Air Force Base, TX: Air Force Systems Command.
Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods, 4*, 192–211.
Lohman, D. F. (1979). *Spatial ability: A review and reanalysis of the correlational literature (Tech. rep. No. 8).* Stanford, CA: Stanford University, School of Education.
Lord, F. M. (1956). A study of speed factors in tests and academic grades. *Psychometrika, 21*(1), 31–50.
Maruyama, M. (1998). *Basics of structural equation modeling.* Thousand Oaks, CA: Sage.
McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive *Gf-Gc* framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.). *Contemporary intellectual assessment: Theories, tests and issues* (pp. 151–179). NY: Guilford Press.
McNemar, Q. (1940). A critical examination of the University of Iowa studies of environmental influences upon the IQ. *Psychological Bulletin, 37*(2), 63–92.
Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2002). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*, 49–100.
Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*(4), 379–387.
Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods, 10*(2), 178–192.
Raven, J. C., Court, J. H., & Raven, J. (1977). *Raven's progressive matrices and vocabulary scales.* New York: Psychological Corporation.
Redick, T. S., Broadway, J. M., Meier, M. E., et al. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment, 28*(3), 164–171.
Ruthsatz, J., Detterman, D., Griscom, W. S., & Cirullo, B. A. (2008). Becoming an expert in the musical domain: It takes more than just practice. *Intelligence, 36*, 330–338.
Salthouse, T. A., Pink, J. E., & Tucker-Drob, E. M. (2008). Contextual analysis of fluid intelligence. *Intelligence, 36*, 464–486.
Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350–353.
Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication.* Urbana: The University of Illinois Press.
Siedlecki, K. L., & Salthouse, T. A. (2014). Using contextual analysis to investigate the nature of spatial memory. *Psychonomic Bulletin & Review, 21*, 721–727.
Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120.
Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. (2016). Do "brain training" programs work? *Psychological Science in the Public Interest, 17*, 103–186.
Spearman, C. (1904). "general intelligence," objectively determined and measured. *American Journal of Psychology, 15*, 201–293.
Spearman, C. (1927). *The nature of "intelligence" and the principles of cognition.* New York: MacMillan.
Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied, 74*, 1–29 (11, Whole No. 498).
Sternberg, R. J. (1986). Haste makes waste versus a stitch in time? A reply to Vernon, Nador, and Kantor. *Intelligence, 10*(3), 265–270.
Thompson, G. (1919). On the cause of hierarchical order among correlation coefficients. *Proceedings of the Royal Society, A. 95. Proceedings of the Royal Society, A* (pp. 400–408).
Thorndike, R. L. (1947). *Research problems and techniques (report no. 3). Army air forces aviation psychology program research reports #3.* Washington DC: U.S. Government Printing Office.
Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques.* New York: John Wiley & Sons.
University of Illinois. https://admissions.illinois.edu/Apply/Freshman/profile Retrieved from the Internet on January 14, 2020.
Van Dyke, J. A., Johns, C. L., & Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition, 131*, 373–403.
Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations: A group test of three-dimensional. Spatial visualization. *Perceptual and Motor Skills, 47*, 599–604.
Vernon, P. E. (1950). *The structure of human abilities.* New York: John Wiley & Sons.
Vernon, P. E., & Parry, J. B. (1949). *Personnel selection in the British forces.* London: University of London Press.
Vernon, P. A., Nador, S., & Kantor, L. (1985). Reaction times and speed-of-processing: Their relationship to timed and untimed measures of intelligence. *Intelligence, 9*(4), 357–374.
Wechsler, D. (1975). Intelligence defined and undefined. *American Psychologist, 30*(2), 135–139.
Wellman, B. L. (1940). Iowa studies on the effects of schooling. In G. M. Whipple (Ed.). *The thirty-ninth yearbook of the National society for the study of education: Intelligence: Its nature and nurture. Part II. Original studies and experiments* (pp. 377–399). Bloomington, IL: Public School Publishing Co.
Wiberg, M., & Sunström, A. (2009). A comparison of two approaches to correction of restriction of range in correlation analysis. *Practical Assessment, Research & Evaluation, 14*(5), 1–9.
Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41*, 67–85.
Wilkinson, L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604.
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, D.C.: American Psychological Association.
Wonderlic, E. F. (1981). *Wonderlic personnel test manual.* Northfield, IL: E. F. Wonderlic & Associates.
Woodrow, H. (1946). The ability to learn. *Psychological Review, 53*, 147–158.

Phillip L. Ackerman[a,*], David Z. Hambrick[b]
[a] *Georgia Institute of Technology, USA*
[b] *Michigan State University, USA*
E-mail address: plackerman@gatech.edu (P.L. Ackerman).

* Corresponding author at: School of Psychology, Georgia Institute of Technology, 654 Cherry Street, MC 0170, Atlanta, GA 30332-0170, USA.