5-6-2015

Building a Better IQ: g Loadings of IQs Experimentally Controlled for Subtest Number, Heterogeneity, g Loading Saturation, and Weighting

Ryan Lee Farmer

BUILDING A BETTER IQ: *G* LOADINGS OF IQS EXPERIMENTALLY CONTROLLED FOR SUBTEST NUMBER, HETEROGENEITY, *G* LOADING SATURATION, AND WEIGHTING

by

Ryan L. Farmer

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Psychology

The University of Memphis

August, 2015

Abstract

Farmer, Ryan Lee. Ph.D. The University of Memphis, August, 2015. Building a Better IQ*:* *g* Loadings of Experimentally Controlled IQs for Subtest Number, Heterogeneity, *g* Loading Saturation, and Weighting. Major Professor: Randy G. Floyd.

Intelligence tests produce IQs which are interpreted as indexes of psychometric *g*. It is thus important to promote the accuracy of those IQs, and to isolate the characteristics of IQs that result in more and less accuracy. In this study, we identified a number of characteristics of IQs that vary across tests, including subtest number, subtest selection methods (e.g., selecting based on heterogeneity, subtest *g* loading, and combined approaches), and weighting procedures. We created experimentally manipulated IQs to systematically test the influence of these characteristics on IQ *g* loadings using three archival datasets. Using cross-battery confirmatory analysis, *g* loadings for each experimental IQ were calculated. Results indicate that increasing subtest number and selecting subtests based on their *g* loading produce IQs with the highest *g* loadings while other methods such as heterogeneous subtest selection and weighting had a nonsignificant influence on IQ *g* loading. The final model accounted for approximately 46% of the variance in IQs that is attributable to psychometric *g*. Discussion regarding directions for future research and implications for test selection and development is provided.

Key Words, Intelligence; composite reliability; *g* loading; psychometric *g*; IQ; construct validity.

Table of Contents

List of Tables

Building a Better IQ*:* *g* Loadings of Experimentally Controlled IQs for Subtest Number, Heterogeneity, *g* Loading Saturation, and Weighting

IQs continue to be useful in applied psychological practice due to diagnostic and taxonomical guidelines (American Association on Intellectual and Developmental Disabilities [AAIDD], 2010; American Psychological Association [APA]. 2000; World Health Organization [WHO], 2010), federal legislation (Individuals with Disabilities Education Improvement Act [IDEIA], 2004), and for predicting important social outcomes such as academic achievement and employment (Schmidt & Hunter, 2004; Sternberg, Grigorenko, & Bundy, 2001; von Stumm, Hell, & Chamorro-Premuzic, 2011). Facilitating its continued use across fields of psychology, IQs are known to have incomparable psychometric properties, with internal consistency reliability coefficients that are frequently above .95 (Kranzler & Floyd, 2013). As a result of these excellent psychometric properties, some researchers in the area of intelligence test psychometrics have argued that test interpretation should be based primarily on the produced IQ (Canivez, 2013; Watkins, Glutting, & Lei, 2007; Kranzler & Floyd, 2013).

The source of individual differences across all cognitive tasks is termed psychometric *g*. Based on an extensive analysis of more than 460 data sets, Carroll (1993) theorized that there were three levels of abilities that varied according to their generality. In this model, psychometric *g* is positioned at the apical and most general level, stratum III. Moving from high to low generality, measures consisting of similar content (e.g., language-based content) or processes (e.g., short-term memory) compose

Carroll's stratum II, and narrowly focused abilities (e.g., analogies) and skills (e.g., recalling numbers in sequence) compose stratum I.

Intelligence test developers maintain Carroll's (1993) structure, in part, by providing stratum III composite scores (i.e., IQs) for interpretation. Several studies have now shown that the calculated psychometric *g* represented in stratum III is either perfectly correlated or nearly perfectly correlated across intelligence tests (Floyd, Reynolds, Farmer, & Kranzler, 2013; Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004; Johnson, te Nijenhuis, & Bouchard, 2008). These findings indicate that although the engineering of the test may vary, the underlying source of individual variances across mental ability tasks remains the same.

***IQ g* Loadings**

It has long been posited that IQs are a strong, but not perfect, stand-in for psychometric *g*, with estimations for IQ to *g* correlations at around .80 (Jensen, 1998). The effect of psychometric *g* on any variable is measured as a *g loading*, a standardized coefficient with a hypothetical range from .00 (indicating no relation) to 1.00 (indicating a perfect relation). Interpretation guidelines have indicated that *g* loadings for any intelligence test of .70 or higher are considered to be strong (Floyd, McGrew, Barry, Rafael, & Rogers, 2009; McGrew & Flanagan, 1998). Recent research has shown that these estimates for IQs were conservative. Using cross-battery confirmatory factor analysis, wherein a derived psychometric *g* from one intelligence test is compared to an IQ from a second intelligence test, Farmer et al. (2013) found across five full-length intelligence tests uncorrected IQ *g* loadings ranging between .84 and .93, with an average of .90. Four of the six IQs evaluated met or exceeded the .90 threshold for interpretation.

Alternately, the percentage of variance in any variable that is attributable to *g* can be referred to as *g saturation.* It is the squared *g* loading in most cases, and it has a hypothetical range from 0% (indicating no variation attributable to *g*) to 100% (indicating all variation attributable to *g*). Reynolds, Floyd, and Niileksala (2013) investigated similar questions to Farmer et al. (2013) using omega coefficients in tandem with confirmatory factor analysis (CFA) across three intelligence tests, estimating the amount of variance in IQs explained by *g* and stratum II abilities. Reynolds et al. found that across the three full-length intelligence tests, IQ *g* saturation for all age ranges ranged between 82% and 83%. These values equate to IQ *g* loadings between .90 and .91.

“Vehicles of *g*”

When engineers are designing vehicles for consumers, a number of considerations must be made regarding various characteristics of the vehicle: weight, horsepower, fuel consumption, speed, and even amenities. With all of the variables at play, not all vehicles are made equally; some have better fuel efficiency, whereas others are capable of higher speeds or towing capacity. Some vehicles are more fuel efficient but lack additional features such as cargo space. In this same sense, intelligence tests vary across test developers and often have strengths and weaknesses depending upon the characteristics of the test itself. Some intelligence tests provide a wider sampling of stratum II abilities, whereas others may feature expedited administration times. Jensen (1998) discussed “vehicles of *g*” (p. 309). Vehicles do not have to look alike or even share similar content to adequately convey *g*, but the goal is still the same: to accurately measure an individuals’ level of *g*. The question at hand is which characteristics of IQs contribute most to the accuracy with which we measure the *g* factor.

At least five methods have been employed by test developers to create IQs. These methods include ensuring heterogeneity of subtests contributing to the IQ, increasing the number of subtests contributing to the IQ, using only highly *g*-loaded subtests in the creation of the IQ, and weighting of subtest scores contributing to the IQ. These design characteristics are not mutually exclusive; in particular, heterogeneity requires a sizable number of subtests. Additionally, many tests use multiple approaches in building their IQs. For example, the Woodcock–Johnson III Tests of Cognitive Abilities (WJ III COG; McGrew, & Woodcock, 2001) General Intellectual Ability-Standard (GIA-Standard) and General Intellectual Ability-Extended (GIA-Extended) were both formed from tests designed to measure different stratum II abilities. Additionally, contributing subtests were weighted in order to more precisely measure *g*. The sections that follow address each of these characteristics more thoroughly and their use in some of the most prominent individually administered intelligence tests are presented as examples.

Heterogeneity

It has been argued that intelligence tests best measure *g* when they sample a wide range of cognitive abilities at varying strata. Humphreys (1979) argued that appropriate sampling of the *g* factor required “a large number of items heterogeneous in content” (p. 115). Expanding upon this assertion, Jensen (1998) suggested that sampling from a variety of highly *g*-loaded and diverse subtests was the optimal approach for measuring the *g* factor. This diversity of content is not problematic; in fact, the aggregation of multiple heterogeneous scores seems to lead to higher *g* loadings because specific variance associated with lower-strata abilities and individual subtests are averaged out whereas that variance attributable to the *g* factor accrues (Maynard, Floyd, Acklie, &

Houston, 2011; Rushton, Brainerd, & Pressley, 1983). In this vein, Gustafsson (2002) argued that, as constructs become more general (such as *g*), measurement of those constructs must become more heterogeneous. As such, using heterogeneous subtests results in an IQ that is a more precise vehicle of *g*.

The WJ III COG (McGew & Woodcock, 2001) is a intelligence test that utilizes heterogeneity to measure *g*. One of its IQs, the GIA-Standard, is composed of seven subtests, each of which is designed to measure a different stratum II ability: Auditory Processing (Ga), Long-Term Retrieval (Glr), Visual Processing (Gv), Crystallized Intelligence (Gc), Processing Speed (Gs), Fluid Intelligence (Gf), and Short-Term Memory (Gsm). The second of its IQs, the GIA-Extended is composed of 14 subtests equally sampling from seven stratum II abilities. Intelligence tests such as the Differential Ability Scales, Second Edition (DAS-II; Elliott, 2007) and the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV; Wechsler, 2003) feature IQs composed of three and five different stratum II abilities (Chen, Chen, Keith, & Chang, 2009; Keith, Fine, Reynolds, Taub, & Kranzler, 2006; Keith, Quirk, Schartzer, & Elliott, 2009), respectively; it is clear that the authors of the WJ III COG made substantial efforts to diversely sample from the spectrum of cognitive abilities.

The Aggregation Hypothesis

Rushton et al. (1983) defined the principle of aggregation, stating that a greater number of measurements is always a more precise and stable indicator of a construct than would be a single measurement. Perhaps a holdover from classical true-score theory (i.e., classical test theory), which assumes that infinite observations would be equivalent to an individual’s true score (Allen & Yen, 2002), a common assumption in intelligence testing is that more is better in terms of subtests. The availability of brief intelligence tests and abbreviated IQs from full-length batteries, as well as full-length tests employing a relatively small number of subtests (e.g., the Reynolds Intellectual Assessment Scales; Reynolds & Kamphaus, 2003), has brought professional attention to the issue of whether or not such tests are appropriate for diagnostic assessment (Homack & Reynolds, 2007). However, the impact of subtest number on IQ-g relations has never been directly assessed.

Intelligence tests vary greatly regarding the number of subtests that contribute to their IQ. Tests such as the WJ III COG (McGrew & Woodcock, 2001) and the WISC-IV (Wechsler, 2003) are two examples that feature a relatively large number of subtests used in IQ construction. As noted previously, the WJ III COG GIA-Extended stems from 14 subtests, and the WISC-IV Full Scale IQ (FSIQ) stems from 10 subtests. In contrast, IQs such as the WJ III COG GIA-Standard (McGrew & Woodcock, 2001), the DAS-II General Conceptual Ability (GCA; Elliott, 2007), and the RIAS Composite Intelligence Index (CIX; Reynolds & Kamphaus, 2003) stem from 7, 6, and 4 subtests, respectively. IQs from brief intelligence tests and abbreviated IQs from full-length batteries sometimes feature even fewer subtests (Kranzler & Floyd, 2013). For example, the Wechsler Abbreviated Scales of Intelligence, Second Edition (WASI; Wechsler, 2011), a brief intelligence test, yields IQs from either 4 or 2 subtests, whereas the WJ III COG (McGrew & Woodcock, 2001) Brief Intellectual Ability score, an abbreviated IQ, is yielded from 3 subtests. Across IQs and abbreviated IQs from brief and full-length batteries, the number of subtests that contribute to an IQ varies greatly, from as few as 2 to as many as 14.

The aggregation hypothesis (Gustafsson, 2002; Rushton et al., 1983) posits that as the number of varied measures increases, the reliability and construct validity of composite scores (e.g., IQs) will also increase. It is unclear whether it is purely the number of items, as argued by classical true-score theory (Allen & Yen, 2002), the impact of selecting heterogeneous measures (Humprheys, 1979; Jensen, 1998), or the interaction of these methods that results in increased reliability and construct validity.

Highly *g*-Loaded Subtests

Researchers such as Gottfredson (1998, 2003) and Jensen (1998) have stated that different academic and vocational tasks employ differing levels of *cognitive complexity*, which is defined as the quality or state of being hard to mentally separate, analyze, or solve. A cognitively complex task is one that requires more "mental manipulations" in order to complete it successfully (te Nijenhuis, van der Flier, 2005). Additionally, tasks with higher levels of cognitive complexity tend to yield higher *g* loadings, and are therefore better—perhaps more pure?—measures of psychometric *g* than are tests that are less complex (Jensen, 1998; Humphreys, 1979).

Studies, such as those by Taub and McGrew (2004) and Carroll's (1993) seminal work, have repeatedly shown that measures of the stratum II ability Processing Speed (e.g., the WISC-IV Coding subtest) are less cognitively complex than those measuring the stratum II ability Fluid Intelligence (e.g., the WISC-IV Matrix Reasoning subtest). Jensen (1998) specifically argued that we should use "substantially *g*-loaded tests" (p. 309) in the development of IQs. There appear to be no peer-reviewed articles or book chapters that have addressed the effects of limiting IQ composition to subtests with only high *g* loadings. Related, by excluding subtests with lower *g* loadings, test developers would be inadvertently restricting the range of cognitive abilities they have available for aggregation into their IQ score, which would counter—albeit in a limited way—the argument for ensuring heterogeneity of IQ composition (which was previously addressed).

The DAS-II (Elliott, 2007) emphasizes the importance of using subtests with high *g*-loadings in IQ composition. Elliott (2012) added that, regarding DAS-II test structure, "the core subtests are relatively strongly *g*-related and therefore measure complex processing and conceptual ability" (p. 339). Specifically, the DAS-II (Elliott, 2007) GCA composite score was developed to "obtain a valid and focused measure of a central component of intellectual ability" (p. 17). Subtests with traditionally lower *g*-loadings (i.e., those belonging to stratum II abilities Processing Speed or Working Memory) are relegated to diagnostic subtests and are available but not included in the GCA. As such, the GCA is composed only of those subtests with high *g* loadings. Maynard and colleagues (2011) found that the core IQ were more highly *g*-loaded than were those subtests included for diagnostic purposes.

Weighting

The weighting of scores to increase the influence of some scores above others is a common procedure both in psychometric and applied settings. Arguably, the logic is that scores stemming from more representative tasks (e.g., semester-long projects) should have more of an impact on overall grades in a course than should scores stemming from less representative tasks (e.g., a single homework assignment). Similar to the logic employed by Elliott (2007) in the development of the DAS-II GCA, weighting allows for heterogeneous sampling and consideration of *g* loadings when selecting subtests to contribute to IQs. Jensen (1998) argued that “optimally weighted composite” scores are better representations of psychometric *g*; however, Jensen added that, for practical applications, the increase in accuracy is insignificant. Similar to the other characteristics addressed, there appear to be no peer-reviewed articles or book chapters that have addressed the effects of weighting subtest scores on IQ *g* loadings.

The WJ III COG (McGrew & Woodcock, 2001) is the only intelligence test to use weighted subtest scores in the composition of its IQs, the General Intellectual Ability – Standard and the General Intellectual Ability - Extended scores. McGrew and colleagues (2001) use a sophisticated procedure for determining the optimal subtest weights, utilizing principal component analysis and polynomial curve-fitting procedures, to most accurately account for the variance in a collection of tests (Schrank, McGrew, & Woodcock, 2001), and to provide the most theoretically pure measure of *g*. In contrast, tests such as the WISC-IV (Wechsler, 2003), DAS-II (Elliott, 2007), and others weight all subtests equally in their contribution to IQs.

Purpose of the Study and Hypotheses

Assessing the IQs yielded by current intelligence tests (cf. Farmer, Floyd, Reynolds, & Kranzler, 2013; Reynolds et al., 2013) are insufficient to understand what leads some IQs to have higher *g* loadings than others. Only by systematically creating IQs with various specifications can the effects of key influences on IQ *g* loadings be evaluated. The purpose of the study was to examine how an IQ’s *g* loading can be influenced by the methods that are used to create it. In particular, this study was designed to investigate the influence of (a) varying the number of subtests, (b) the heterogeneity of the subtests, (c) selecting subtests based on their *g* loading magnitude, and (d) weighting subtest scores. Understanding how these characteristics affect IQ *g* loadings informs both practice and test development. By understanding which characteristics provide a more precise vehicle of *g* and by default, better predictive power, it becomes possible to optimally select intelligence tests in practice that maximize these components. With this understanding, test developers will be better able to develop IQs that more accurately measure *g* in future editions of intelligence tests.

It was hypothesized that, based on prior research, IQ *g* loadings would undergo a negligible increase as the number of contributing subtests increase beyond four. Second, it was hypothesized that as the heterogeneity of content contributing to the IQ increased, IQ *g* loading would increase, as predicted by the aggregation hypothesis (Gustafson, 2002; Rushton et al., 1983). Third, it was hypothesized that using subtests with higher *g* loadings would result in an IQ that more accurately reflects *g*, resulting in an increase in *g* loading. Fourth, as Jensen (1998) argued, it was hypothesized that weighting of subtests scores would have a minimal impact on IQ *g* loadings. Finally, it was hypothesized that the interaction between subtest number and heterogeneity would have the greatest positive influence on IQ *g* loadings.

Method

This study will report the results of analyses of data sets from three samples drawn during the development of the DAS-II (Elliot, 2007) and the WJ III (McGrew & Woodcock., 2001).

Participants

Sample 1. Elliot (2007), Floyd et al. (2013), and Farmer et al. (2013) described data on 200 children ages 6 to 16 years, 11 months who completed both the DAS-II (Elliot, 2007) and then the WISC-IV (Wechsler, 2003). Of those sampled, 100 (50%) were girls. In addition, 50 (25%) were African American, 13 (6.5%) were Asian, 54 (27%) were Hispanic, 70 (35%) were White, and 13 (6.5%) were listed as Other (Elliot, 2007).

Sample 2. McGrew and Woodcock (2001), Phelps, McGrew, Knopik, and Ford (2005), Floyd, Clark, and Shadish (2008), Floyd et al. (2013), and Farmer et al. (2013) described data on 150 randomly selected children ages 8 to 12 years, 4 months who completed the WJ III COG (McGrew & Woodcock, 2001) and the WISC-III (Wechsler, 1991) in counterbalanced order (McGrew & Woodcock, 2001). Of those sampled, 66 were girls (44%). Additionally, 2 (1.3%) were African American and 148 (98.7%) were White (McGrew & Woodcock, 2001; Phelps, McGrew, Knopik, & Ford, 2005).

Sample 3. McGrew and Woodcock (2001), Sanders, McIntosh, Dunham, Rothlisberg, and Finch (2007), Floyd et al., (2008), Floyd et al. (2013), and Farmer et al. (2013) described data on 135 randomly selected children ages 8 to 13 who completed the Differential Ability Scales (DAS, Elliot, 1990) and the WJ III Tests of Cognitive Abilities (McGrew & Woodcock, 2001) in counterbalanced order (McGrew & Woodcock, 2001). Of those sampled, 69 (51.9%) were girls. In addition, 6 (4.5%) were African American and 127 (95.5%) were White (McGrew & Woodcock, 2001).

Measures

Differential Ability Scales, Second Edition. The study would include data from 10 DAS-II (Elliot, 2007) subtests: Matrices, Pattern Construction, Recall of Designs, Rapid Naming, Recall of Digits Backward, Recall of Sequential Order, Sequential and Quantitative Reasoning, Speed of Information Processing, Verbal Similarities, and Word Definitions. Subtest scores had median reliability coefficients equal to or above .78 across ages 6 to 17 in the norming sample. Subtest information, including means and standard deviations from Sample 1, internal consistency reliability coefficients from the standardization sample, stratum II ability classifications, and *g* loadings from Floyd et al. (2013) are presented in Table 1.

Wechsler Intelligence Scale for Children, Fourth Edition. The study included data from 14 WISC-IV (Wechsler, 2003) subtests: Arithmetic, Block Design, Cancellation, Coding, Comprehension, Digit Span, Information, Letter-Number Sequencing, Matrix Reasoning, Picture Completion, Picture Concepts, Similarities, Symbol Search, and Vocabulary. Subtests had median reliability coefficients equal to or above .79 across ages 6 to 16 in the norming sample. Subtest information, including means and standard deviations from Sample 1, internal consistency reliability coefficients from the standardization sample, stratum II ability classifications, and *g* loadings from Floyd et al. (2013) are presented in Table 1.

Table 1

Sample 1 Subtest Characteristics

Test and subtest	*M*	*SD*	Median IRC	Stratum II Grouping	*g* loading
DAS-II					
Digits Backward	49.31	8.60	.90	Gsm	.54
Matrices	49.70	8.63	.86	Gf	.72
Pattern Construction	49.64	7.79	.96	Gv	.67
Rapid Naming	49.90	8.58	.81	Gs	.33
Recall of Designs	51.07	8.51	.85	Gv	.54
Recall of Sequential Order	49.17	8.16	.93	Gsm	.59
Sequential & Qualitative Reasoning	50.07	8.17	.93	Gf	.74
Speed of Information Processing	50.80	8.70	.92	Gs	.37
Verbal Similarities	50.55	7.70	.78	Gc	.50
Word Definitions	49.48	8.39	.83	Gc	.48
WISC-IV					
Arithmetic	10.18	2.82	.86	Gsm	.76
Block Design	10.79	2.86	.87	Gv	.65
Cancellation	9.49	3.01	.79	Gs	.29
Coding	9.60	2.97	.87	Gs	.45
Comprehension	10.16	2.81	.81	Gc	.56
Digit Span	10.38	2.87	.88	Gsm	.50
Information	10.11	2.62	.86	Gc	.64
Letter–Number Sequencing	10.00	2.64	.90	Gsm	.63
Matrix Reasoning	11.08	2.87	.89	Gf	.72
Picture Completion	10.11	2.99	.84	Gv	.52
Picture Concepts	11.07	2.69	.82	Gf	.54
Similarities	10.48	2.47	.86	Gc	.66
Symbol Search	10.37	2.59	.80	Gs	.44
Vocabulary	10.30	2.55	.90	Gc	.72
Word Reasoning	10.57	2.75	.80	Gc	.58

Note. WISC-IV = Wechsler Intelligence Scales for Children, Fourth Edition. DAS-II = Differential Ability Scales, Second Edition. Gc = Comprehension–Knowledge, Gsm = Short-Term Memory, Gs = Processing Speed, Gv = Visual Processing, Gf = Fluid Reasoning, and IRC = Internal Reliability Coefficient.
Means and standard deviations were extracted from analysis employing maximum likelihood estimation, whereas median internal consistency reliability data were obtained from standardization sample data and reported in test manuals. Subtest *g* loadings were calculated by multiplying the first-order factor loading by the second-order factor loading; these calculations will be based on findings from Floyd and colleagues (in press).

Wechsler Intelligence Scale for Children, Third Edition. The study included data from 12 WISC-III (Wechsler, 1991) subtests: Arithmetic, Block Design, Coding, Comprehension, Digit Span, Information, Object Assembly, Picture Arrangement, Picture Completion, Similarities, Symbol Search, and Vocabulary. All subtests, excluding Object Assembly (.68) and Picture Arrangement (.72), had median reliability coefficients above .75 across ages 8 to 13 in the norming sample. Subtest information, including means and standard deviations from Sample 2, internal consistency reliability coefficients from the standardization sample, stratum II ability classifications, and *g* loadings from Floyd et al. (2013) are presented in Table 2.

Woodcock–Johnson III Tests of Cognitive Abilities, Third Edition. The study included data from 13 WJ III COG (McGrew & Woodcock, 2001) subtests: Analysis Synthesis, Auditory Attention, Auditory Working Memory, Concept Formation, Decision Speed, Incomplete Words, Memory for Words, Numbers Reversed, Picture Recognition, Rapid Picture Naming, Retrieval Fluency, Sound Blending, Spatial Relations, Verbal Comprehension, Visual Auditory Learning, and Visual Matching. The subtests had median reliability coefficients equal to or above .70 across ages 8 to 13. Subtest information, including means and standard deviations from Sample 2, internal consistency reliability coefficients from the standardization sample, stratum II ability classifications, and *g* loadings from Floyd et al. (2013) are presented in Table 2. Furthermore, subtest information, including means and standard deviations from Sample 3, internal consistency reliability coefficients from the standardization sample, stratum II ability classifications, and *g* loadings from Floyd et al. (2013) are presented in Table 3.

Table 2

Sample 2 Subtest Characteristics

Test and subtest	*M*	*SD*	Median IRC	Stratum II Grouping	*g* loading
WISC-III					
Arithmetic	10.99	2.88	.78	Gsm	.58
Block Design	11.02	3.39	.85	Gv	.57
Coding	11.29	2.68	.80†	Gs	.37
Comprehension	11.20	3.20	.79†	Gc	.49
Digit Span	11.09	3.02	.84	Gsm	.53
Information	11.95	2.51	.85	Gc	.58
Object Assembly	10.18	2.75	.68	Gv	.45
Picture Arrangement	10.67	3.56	.72	Gv	.28
Picture Completion	10.36	2.85	.76	Gv	.35
Similarities	11.59	2.75	.82	Gc	.61
Symbol Search	11.95	3.30	.76	Gs	.38
Vocabulary	10.56	3.20	.88	Gc	.65
WJ III					
Analysis–Synthesis	105.60	11.43	.89	Gf	.61
Auditory Attention	100.39	12.70	.88	Ga	.20
Auditory Working Memory	102.69	11.87	.87	Gsm	.50
Concept Formation	101.87	10.79	.94	Gf	.66
Decision Speed	102.58	13.51	.85	Gs	.32
Incomplete Words	96.82	16.19	.77	Ga	.33
Memory for Words	101.26	13.59	.77	Gsm	.45
Numbers Reversed	103.51	12.35	.86	Gsm	.47
Picture Recognition	101.45	12.78	.70	Gv	.04
Rapid Picture Naming	100.28	10.65	.97	Gs	.19
Retrieval Fluency	100.97	12.57	.81	Glr	.19
Sound Blending	98.93	13.79	.84	Ga	.32
Spatial Relations	100.22	14.19	.81	Gv	.36
Verbal Comprehension	107.19	12.96	.90	Gc*	.72
Visual–Auditory Learning	102.75	12.78	.86	Glr	.59
Visual Matching	103.19	14.62	.89	Gs	.41

Note. WISC-III = Wechsler Intelligence Scales for Children, Third Edition; WJ III = Woodcock–Johnson III Test of Cognitive Abilities. Gc = Comprehension–Knowledge, Gsm = Short-Term Memory, Gs = Processing Speed, Gv = Visual Processing, Gf = Fluid Reasoning, and IRC = Internal Reliability Coefficient. Means and standard deviations were extracted from analysis employing maximum likelihood estimation, whereas median internal consistency reliability data were obtained from test manuals. Subtest *g* loadings were calculated by multiplying the first-order factor loading by the second-order factor loading; these calculations will be based on findings from Floyd and colleagues (in press).

* The subtest was regressed directly on to the general factor because no other subtests targeting the same stratum II ability were available in the data set. However, the stratum II ability label is based on validity evidence presented in McGrew & Woodcock (2001).

† For Coding and Symbol Search, reliability coefficients are based on raw-score test–retest correlations; these correlations were then corrected for the variability of the appropriate standardization group (Wechsler, 1991).

Differential Ability Scales. The study included data from eight DAS (Elliott, 1990) subtests: Matrices, Pattern Construction, Recall of Designs, Recall of Objects-Immediate, Sequential and Qualitative Reasoning, Similarities, Speed of Information Processing, and Word Definitions. Subtest scores had median reliability coefficients[1] above .75 across ages 8 to 12 in the norming sample. Subtest information, including means and standard deviations from Sample 3, internal consistency reliability coefficients from the standardization sample, stratum II ability classifications, and *g* loadings from Floyd et al. (2013) are presented in Table 3.

[1] We refer to reliability coefficients in general but recognize that there is variation in how these coefficients were obtained. Subtest reliabilities yielded from analysis of norming data stem from internal consistency reliability analysis, for *power* tests, and from test-retest analysis, for *speed* tests (such as those measuring Processing Speed).

Table 3

Sample 3 Subtest Characteristics

Test and subtest	*M*	*SD*	Median IRC	Stratum II Grouping	*g* loading
DAS					
Matrices	54.36	8.88	.85	Gf	.62
Pattern Construction	55.17	8.37	.90	Gv	.47
Recall of Designs	54.12	11.35	.82	Gv	.50
Recall of Objects-Immediate	52.58	8.63	.76	Glr*	.32
Sequential and Qualitative Reasoning	57.41	9.24	.85	Gf	.80
Similarities	54.78	10.73	.79	Gc	.67
Speed of Information Processing	54.97	9.88	.91	Gs*	.37
Word Definitions	53.13	10.18	.84	Gc	.60
WJ III					
Analysis–Synthesis	109.20	13.99	.89	Gf	.72
Auditory Attention	108.04	15.33	.88	Ga	.30
Auditory Working Memory	104.33	15.85	.87	Gsm	.59
Concept Formation	106.38	13.90	.94	Gf	.67
Decision Speed	105.77	15.82	.85	Gs*	.43
Incomplete Words	102.97	13.92	.77	Ga	.48
Memory for Words	104.91	14.18	.77	Gsm	.48
Numbers Reversed	105.08	16.29	.86	Gsm	.39
Picture Recognition	93.26	13.36	.70	Gv	.29
Sound Blending	101.91	13.87	.84	Ga	.48
Spatial Relations	98.71	17.03	.81	Gv	.19
Verbal Comprehension*	109.51	13.95	.90	Gc*	.75
Visual–Auditory Learning*	104.51	13.79	.86	Glr*	.53

Note. DAS = Differential Ability Scales; WJ III = Woodcock–Johnson III Test of Cognitive Abilities. Gc = Comprehension–Knowledge, Gsm = Short-Term Memory, Gs = Processing Speed, Gv = Visual Processing, Gf = Fluid Reasoning, Glr = Long-Term Storage and Retrieval, Ga = Auditory Processing, and IRC = Internal Reliability Coefficient.

Means and standard deviations were extracted from analysis employing maximum likelihood estimation, whereas median internal consistency reliability data were obtained from test manuals. Subtest *g* loadings were calculated by multiplying the first-order factor loading by the second-order factor loading; these calculations will be based on findings from Floyd and colleagues (in press).

* Subtests were regressed directly on to the general factor because no other subtests targeting the same stratum II ability were available in the data set. However, the stratum II ability label is based on validity evidence presented in Elliott (2007) and Keith, Quirk, Schartzer, and Elliott (2009) for the DAS. and evidence presented in McGrew & Woodcock (2001) for the WJ III

Analysis

General methods of composite formulation. In order to test the influence of the test characteristics on the *g* loading of IQs, several methods were used to generate composite scores stemming from intelligence test subtest scores. The methods used to create these experimental IQs are described in more detail in the sections that follow.

Number of subtests. The first manipulated characteristic is the number of subtests contributing to the experimental IQ. IQs were created from 2, 3, 4, 7, 10, and 15 subtests, depending upon the availability of subtests within the sample. These values of 2, 3, 4, 7, 10, and 15 were selected based on a review of the common number of subtests contributing to IQ and abbreviated IQs from brief and full-length batteries (Kranzler & Floyd, 2013). This characteristic is overarching, and all other characteristics were created within each of the aforementioned conditions. See the first column in Table 4.

Heterogeneous stratum II ability sampling. In order to experimentally test the influence of heterogeneity on the *g* loading of IQs, randomly selected subtests from within randomly selected stratum II ability areas were used to create composites. First, subtests that are indicators of stratum II ability areas were identified based on confirmatory factor analysis studies by Bickley, Keith, and Wolfe (1995), Keith, Low, Reynolds, Patel, and Ridley (2010), Sanders, McIntosh, Dunham Rothlisberg, and Finch (2007), Phelps et al. (2005), Keith, Fine, Reynolds, Taub, and Kranzler (2006), and Keith, Kranzler, and Flanagan (2001), as well as test manuals (Elliott, 1990; Elliott, 2007; Wechsler, 1991; Wechsler, 2003; McGrew & Woodcock, 2001). Stratum II ability classifications for each subtest were identified by sample in the fifth column of Tables 1, 2, and 3. Second, a list of stratum II ability areas measured across subtests was created.

Third, a stratum II ability area was randomly selected via assigning numbers to each available ability area and using a random integer generator via random.org (Haahr, 2012). When multiple subtests associated with this stratum II ability area were present, a subtest indicator within that area was randomly selected using the same method as previously described. After the subtest from the first stratum II ability area was selected, additional stratum II areas were selected from all remaining areas. For example, when developing an experimental IQ for Sample 1 from WISC-IV subtests, Crystallized Intelligence was randomly selected first. Based on this result, a subtest that has been shown to be an indicator of Crystallized Intelligence would be chosen at random from other indicators (e.g., Information). The next step would be to select another stratum II ability area from the remaining ability areas measured by the WISC-IV: Fluid Intelligence, Visual Processing, Short-Term Memory, and Processing Speed. This process would then be repeated until the maximum number of subtests for the IQ had been selected. If all ability areas had been sampled from and additional subtests were needed, such as in the case of an IQ stemming from 7 or more subtests, all ability areas would be returned to the selection pool, and the process would be repeated until all subtests had been selected. See the second column in Table 4.

Magnitude of subtest g loading. In order to experimentally test the influence of the magnitude of subtest *g* loadings on the *g* loading of IQs, subtests were rank ordered by their *g* loadings. Subtest *g* loadings were calculated by multiplying the first-order factor loading by the second-order factor loading based on results from the models presented in Floyd and colleagues (2013) and derived from maximum likelihood estimation and the same data sets included in this study. The *g* loadings have been provided for each subtest in the sixth column in Tables 1, 2, and 3. In those cases in which more than one subtest had the same *g* loading, one subtest was randomly selected for use first in the composite using the same method as described previously for randomly selecting stratum II ability areas. For example, in creating an experimental IQ in this condition that stems from 2 subtests from the WISC-IV, the first subtest to be chosen would be Arithmetic with a *g* loading of .76. Both Matrix Reasoning and Vocabulary have a *g* loading of .72; in this case, the random selection of the equivalently *g*-loaded subtests would occur in order to select the next subtest. See the third column in Table 4.

Both heterogeneous ability sampling and magnitude of subtest g loadings. A combination of heterogeneity and *g* loading subtest selection methods was also used to experimentally test for their interaction effect on the *g* loading of IQs. First, heterogeneity subtest selection methods (described in the previous paragraphs) were completed wherein subtests were randomly selected by stratum II ability area. The methodology was altered, however, in order to select subtests when multiple subtests were available within an ability area. Rather than select them randomly, as previously described, the subtest with the highest *g* loading within each ability area was first be selected. For example, in creating an experimental IQ in this condition that stems from 4 subtests from the WISC-IV, a stratum II ability area was first randomly selected (e.g., Crystallized Intelligence); from within that ability area, the subtest with the highest *g* loading (e.g., from Crystallized Intelligence, Vocabulary has the highest *g* loading of .72). The next step is to randomly select a stratum II ability area from the remaining ability areas (as described previously); if Short-Term Memory was selected, then from within that classification, the highest *g* loaded subtest, Arithmetic at .76, would be tapped as the second subtest to contribute to the IQ. This process is repeated until all subtests have been selected. See the fourth column in Table 4.

Weighting. After subtests were selected to form each experimental IQ, the subtests' *g* loadings, as seen in the sixth column of Tables 1, 2, and 3, were used to form weighted composites. More specifically, Bartlett weights (DiStefano et al., 2009), using the *g* loading of subtests, were used to establish scores that are the most quantitatively pure measures of *g*. See the bottom half of Table 4.

Table 4

Example experimental IQ table across all variations.

	Subtest Number	Magnitude of subtest *g* loadings	Heterogeneous Sampling	Combined	Random
Unweighted	2	-	-	-	
	3	-	-	-	
	4	-	-	-	
	7	-	-	-	
	10	-	-	-	
	15	-	-	-	

	Subtest Number	Magnitude of subtest *g* loadings	Heterogeneous Sampling	Combined	Random
Weighted	2	-	-	-	
	3	-	-	-	
	4	-	-	-	
	7	-	-	-	
	10	-	-	-	
	15	-	-	-	

Random. Finally, in order to test the effect of subtest selection method, subtest number, and weighting systematically, subtest selection was randomized. That is, in the random subtest selection method condition, all available subtests from a given intelligence test were entered into a random list generator (Haahr, 2012). The lists were randomized, and composites were created based on these lists, regardless of theory.

Sample-specific methods of composite formulation. The following paragraphs will describe the demographic and collection methodologies of the three archival datasets analyzed.

Sample 1. The first models constructed from sample 1 was a hierarchical model of the DAS-II (Elliott, 2007) with a second-order general factor (see Figure 1); 10 subtests from the DAS-II (Elliott, 2007) were used to model the general factor consistent with prior publications (Floyd et al., in press). Characteristics (i.e., number of subtests, heterogeneity, and subtest *g* loading magnitude) were then be varied to create 40 experimental IQs from 15 subtests from the WISC-IV (Wechsler, 2003). Finally, each experimental IQ was correlated individually with the general factor derived from the DAS-II in order to obtain *g* loadings for each experimental IQ.

Subsequently, the second model constructed from sample 1 was a hierarchical model of the WISC-IV (Wechsler, 2003) with a second-order general factor; 15 subtests from the WISC-IV (Wechsler, 2003) were used to model the *g* factor consistent with prior publications (Floyd et al., 2013). Characteristics were varied to create 44 experimental IQs from 10 subtests from the DAS-II (Elliott, 2007). To form the experimental IQs, norm-referenced subtest scores from the respective test were summed.

Finally, each experimental IQ was correlated individually with the general factor derived from the WISC-IV in order to obtain their *g* loadings.

Sample 2. The first model constructed from sample 2 was a hierarchical model of the WISC-III (Wechsler, 1991) with a second-order general factor; 12 subtests from the WISC-III were used to model the general factor consistent with prior publications (Floyd et al, 2013). Characteristics were varied to create 48 experimental IQs from 16 subtests from the WJ III COG (McGrew & Woodcock, 2001). Finally, each experimental IQ was correlated individually with the general factor derived from the WISC-III in order to obtain *g* loadings for each experimental IQ.

Subsequently, the second model constructed from sample 2 was a hierarchical model of the WJ III COG (McGrew & Woodcock, 2001) with a second-order general factor; 16 subtests from the WJ III COG were used to model the general factor consistent with prior publications (Floyd et al., 2013). Characteristics were then varied to create 40 experimental IQs from 12 subtests from the WISC-III (Wechsler, 1991). To form the experimental IQs, norm-referenced subtest scores from the respective test were summed. Finally, each experimental IQ was correlated individually with the general factor derived from the WJ III COG in order to obtain *g* loadings for each experimental IQ.

Sample 3. The first model constructed from sample 3 was a hierarchical model for the DAS (Elliott, 1990) with a second-order general factor; 8 subtests from the DAS were used to model the general factor consistent with prior publications (Floyd et al, 2013). Characteristics were then varied to create 40 experimental IQs from 13 subtests from the WJ III COG (McGrew & Woodcock, 2001). Finally, each experimental IQ was correlated individually with the general factor derived from the DAS in order to obtain *g* loadings for each experimental IQ.

Subsequently, the second model constructed from sample 3 was a hierarchical model of the WJ III COG (McGrew & Woodcock, 2001) with a second-order general factor; 13 subtests from the WJ III COG were used to model the general factor consistent with prior publications (Floyd et al, 2013). Characteristics were then varied to create 32 experimental IQs from 8 subtests from the DAS (Elliott, 1990). To form the experimental IQs, norm-referenced subtest scores from the respective test were summed. Finally, each experimental IQ was correlated individually with the general factor derived from the WJ III COG in order to obtain *g* loadings for each experimental IQ.

Model structure. Previously, Farmer et al. (2013) utilized cross-battery confirmatory factor analysis to estimate the *g* loading of IQs; the current analysis is an extension of this work. Generally, modeling was conducted using MPlus version 6.1 (L. K. Muthén & B. O. Muthén, 1998-2010) based on models completed in Farmer et al. (2013). Maximum likelihood estimation with robust standard error was used to estimate free parameters, with the assumption that all data are missing at random (Baraldi & Enders, 2010).

The intent was to build models in which the experimental IQs, constructed from subtests from the first intelligence test in each sample, were correlated with a second-order *g* factor that was modeled from the second intelligence test in each sample. As established in Floyd et al. (2013), second-order *g* factors were effectively perfectly correlated between intelligence tests in each sample. The model was established with representative first-order factors (as indicated in Tables 1-3) and a single second-order *g* factor. The next step, as seen in Farmer et al. (2013), was to introduce the experimental IQs to the model, one at a time, and to correlate them with the second-order *g* factor from the second intelligence test, thus estimating the experimental IQs' *g* loadings.

Figure 1 provides an example using the DAS-II (Elliott, 2007) and the WISC-IV (Wechsler, 2003) employed in sample 1. Subtests (e.g., Recall of Designs) from the DAS-II were used to model stratum II and stratum III latent variables. These subtest scores are presented on the left side of the figure and are represented by rectangles. The unique variances and error associated with each subtest are not depicted in Figure 1. First-order factors, representing stratum II ability factors, are represented by ellipses which are drawn immediately to the right of the subtests from which they stem; in this example, stratum II ability areas include Visual Processing (Gv), Crystallized Intelligence (Gc), Processing Speed (Gs), Fluid Intelligence (Gf), and Short-Term Memory (Gsm). Paths are drawn from stratum II ability factors to the subtests used as their indicators; these paths are represented by arrows. For example, the Recall of Designs and Pattern Construction subtests are indicators of the Visual Processing factor.

The unique variances of the stratum II ability factors are represented by circles slightly above the ellipse of their respective ability area, but they are not relevant to this study. A single ellipse, denoted by "*g*," with paths drawn to all stratum II ability areas, represents psychometric *g*. The IQ in this example is a measured variable derived from 10 subtests from the WISC-IV (Wechsler, 2003), and it is thus represented by a rectangle denoted by "IQ." Finally, a double-headed arrow is drawn connecting the stratum III general factor of intelligence and the IQ, representing a correlation.

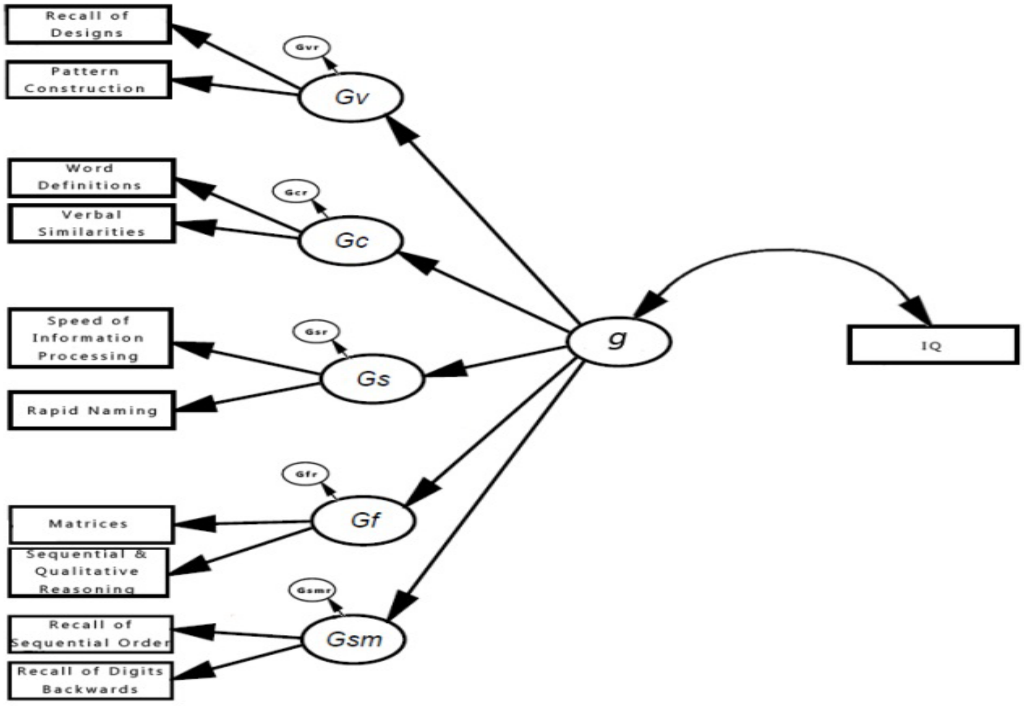

Figure 1. Example model of experimental IQ correlated with *g* established from the Differential Ability Scales, Second Edition, using Cross-Battery Factor Analysis.

Influence of Attributable Factors on IQ *g* Loadings. The *g* loadings of the experimental IQs were compared to one another in order to test the hypothesis that the magnitude of IQ *g* loadings is influenced by the number of subtests, the heterogeneity of utilized subtests, the *g* loading of subtests used, and whether or not those subtests were weighted. Initially, data were entered into a series of sequential regression analyses where the Fisher z-transformed *g* loading (hence forth, *z-g loading*) was used as the dependent variable. Further, to test specific interactions between independent variables and levels of those independent variables, ANOVAs were run.

Results

g loadings

The 244 generated IQ *g* loadings produced through the modeling process varied greatly, with *g* loadings ranging between .488 and .981 (see Appendices A through F), with an average of .824 (*SD* = .099). [2] In addition, mean *g* loadings varied within subtest selection methods as well. In particular, within heterogeneous subtest selection across samples, the average *g* loadings for unweighted IQs ranged between .710 and .878 (*M* = .815, *SD* = .065), whereas the average *g* loadings for weighted IQs ranged between .737 and .907 (*M* = .828, *SD* = .060). Within *g* loading-based subtest selection across samples, the average *g* loadings for unweighted IQs ranged between .862 and .932 (*M* = .891, *SD* = .025), whereas the average *g* loadings for weighted IQs ranged between .839 and .924 (*M* = .884, *SD* = .027). Within the combined subtest selection method across samples, the average *g* loading for unweighted IQs ranged between .784 and .900 (*M* = .843, *SD* = .040), whereas the average *g* loading for weighted IQs ranged between .806 and .902 (*M* = .857, *SD* = .038). Finally, within random subtest selection across samples, the average *g* loading for unweighted IQs ranged between .748 and .872 (*M* = .805, *SD* = .046), whereas the average *g* loaded for weighted IQs ranged between .714 and .871 (*M* = .809, *SD* = .058). These data are presented in Appendix D, Table D1.

Furthermore, mean *g* loadings varied by subtest number across samples and subtest selection methods. In particular, within composites created from two subtests across samples, mean *g* loadings ranged between .692 and .776 (*M* = .741, *SD* = .040).

[2] Averages were calculated by performing a Fisher *z* transformation on independent correlation coefficients, averaging the transformed correlations, and then taking the inverse of the transformation.

Composites created from three subtests across samples resulted in mean *g* loadings ranging between .754 and .829 (*M* = .794, *SD* = .032). Composites created from four subtests across samples resulted in mean *g* loadings ranging between .811 and .853 (*M* = .836, *SD* = .018). Composites created from seven subtests across samples resulted in mean *g* loadings ranging between .846 and .951 (*M* = .899, *SD* = .040). Within composites created from 10 subtests across samples resulted in mean *g* loadings ranging between .864 and .952 (*M* = .909, *SD* = .037). Finally, only the WISC-IV from sample 1 and the WJ III from sample 2 lead to enough IQs formed from 15 subtests to result in an average. The 15-subtest IQs from the WISC-IV in sample 1 had a mean *g* loading of .934 while the 15-subtest IQs from the WJ III in sample 2 had a mean *g* loading of .882. These data are presented in Appendix D, Table D2.

Variables

A number of coding schemes were used in order to effectively test for a variety of effects. First, the categorical variable Sample was coded to reflect from which sample each *z-g* loading was derived. Subtest number was coded as a continuous variable. Further, subtest selection method was initially coded as a single categorical variable (i.e., 1 = heterogeneity, 2 = subtest *g* loading magnitude, 3 = combined, and 4 = random selection). In order to accommodate the sequential regression, this categorical variable was broken into a series of distinct variables representing each selection method coded dichotomously. That is, composites for which subtests were selected based on heterogeneity only were coded “1” for the heterogeneity variable and “0” for all other variables. The combined variable essentially acted as an interaction variable between heterogeneity and subtest *g* loading; only if both were coded as “1,” would the combined variable be coded as “1.” Finally, an interaction term for heterogeneity and subtest number was created by obtaining the product of the two coded variables (i.e., Heterogeneity * SubtestNumber). Additionally, weighting was coded dichotomously to indicate whether a composite used weighting procedures or was a sum of subtest scores.

Overall Model

First, Pearson product moment correlations between dependent and independent variables (see Table 5) were reviewed. A strong statistically significant relation was observed between the *z-g* loadings and Subtest Number. Furthermore, a moderate significant relationship appeared between the *z-g* loadings and the subtest *g* loading magnitude selection method. Neither weighting, the heterogeneity subtest selection method, nor using a combination of heterogeneity and subtest *g* loading magnitude selection methods had a significant relation with the *z-g* loadings.

Table 5

Correlations and Significance of the Fisher Inverse of Composite z-g Loadings with Composite Characteristics

		Subtest g		Subtest	
	Heterogeneity	Loading	Combined	Number	Weighted
z-g loading	-.08 (.23)	.28 (.00)	.03 (.63)	.60 (.00)	.04 (.50)

Note. *z-g* loadings are Fisher z-transformation of *g* loadings. Probability values are presented in parenthesis.

A sequential multiple regression analysis was run wherein the *z-g* loading was regressed on source sample, subtest number, the presence of weighting, and subtest selection method: heterogeneity, *g* magnitude, and combined. First, to ensure that variance associated with the source sample—that is, the archival data set from which the original data were collected—was controlled for in the analysis, sample was entered into the sequential regression. Sample alone resulted in a statistically nonsignificant model that accounted for only 5% of the variance observed in *z-g* loadings, $R^2 = .005$, $F(1,244)=1.321$, $p = .252$. Next, in order to evaluate the effects of subtest number alone, this variable was entered into the regression. Subtest number resulted in a statistically significant increase, allowing the model to predict approximately 38% of the variance observed in composite *z-g* loadings, $\Delta R^2 = .378$, $F(1,243)=149.169$, $p < .001$.

Of greater interest are the results of the third step of the sequential regression. In this step, the construction methods of weighting, selection of subtests by heterogeneity, subtest *g* loading magnitude, and combined approaches, resulted in a statistically significant increase in the variance of composite *z-g* loadings, predicting 48% of total variance, $\Delta R^2=.117$, $F(4,239)=13.991$, $p < .001$. Of the coefficients, only subtest number, subtest selection by *g* magnitude, and subtest selection by heterogeneity and subtest *g* loading magnitude combined resulted in statistically significant contributions to the model ($p < .001$ in all cases); beta weights are provided in Table 6. The use of the combined approach (i.e., heterogeneity and *g* loading magnitude subtest selection) resulted in a negative β, indicating that the use of this method reduced the relation of the resultant composite with *g*. Neither weighting nor subtest selection by heterogeneity resulted in significant contributions to the model.

Table 6

Coefficient Values when z-g Loading is Sequentially Regressed on Sample, Subtest Number, Subtest Selection Method, and Weighting

Model		*B*	SE *B*	β	*B* 95% CI	ΔR^2
1						.005
	Sample	.029	.025	.073	[-.021, .079]	
2						.378*
	Subtest Number	.053	.005	.604	[.046, .063]	
3						.117*
	Heterogeneity	.048	.041	.075	[-.033, .128]	
	Subtest *g* Loading	.286	.042	.450	[.203, .365]	
	Weighting	.018	.029	.028	[-.040, .075]	
	Combined	-.195	.059	-.265	[-.309, .081]	

Note. * Models 2 and 3 were statistically significant at the $p < .001$ level. Model 1 was nonsignificant (p = .252).

Because some of the intelligence tests have an insufficient number of subtests to produce 10- and 15-subtest composites, the previous sequential regression was completed a second time using only composites derived from 7 or fewer subtests. Table 7 presents these results. Once again, sample source was first entered into the regression in an effort to control for differences between the original samples and subtest number was entered second into the regression in an effort to evaluate its effects alone. Sample source alone did not result in a statistically significant model, predicting only 2% of the variance observed in composite *z-g* loadings, $R^2 = .020$, $F(1,190)=3.779$, $p = .053$. During the second step, subtest number was entered into the regression and resulted in a statistically significant model, predicting approximately 41% of the variance observed in composite *z-g* loadings, $\Delta R^2=.399$, $F(1,189)=129.826$, $p < .001$. The second step introduced the construction methods of weighting, selection of subtests by heterogeneity, *g* loading magnitude of subtests, and combined approaches. This step resulted in a statistically significant increase in the variance of composite *z-g* loadings, predicting 59% of total variance, ΔR^2=.189, F(1,185)=22.319, p < .001. Of the coefficients, only subtest number, subtest selection by subtest *g* loading magnitude, and subtest selection by heterogeneity and by subtest *g* loading magnitude combined resulted in statistically significant contributions to the model (p < .001 in all cases); beta weights are provided in Table 7. As in the previous models, the use of the combined approach resulted in a negative β, indicating that the use of this method reduced the relation of the resultant composite with *g*. Once again, neither weighting nor subtest selection by heterogeneity resulted in significant contributions to the model.

Table 7

Coefficient Values when z-g Loading for Composites from 7 or Fewer Subtests is Sequentially Regressed on Sample, Subtest Number, Subtest Selection Method, and Weighting.

Model	*B*	SE *B*	β	*B* 95% CI	ΔR^2
1					.020
Sample	.051	.026	.140	[-.001, .102]	
2					$.399^*$
Subtest Number	.100	.009	.632	[.083, .118]	
3					$.189^*$
Heterogeneity	.041	.039	.069	[-.035, .118]	
Subtest *g* Loading	.334	.039	.563	[.258, .410]	
Weighting	-.001	.027	-.002	[-.055, .053]	
Combined	-.231	.055	-.336	[-.338, -.123]	

Note. *Models 2 and 3 were statistically significant at the p < .001 level. Model 1 was nonsignificant (p = .053).

Subtest Number

One specific purpose of this research was to evaluate the effect of subtest number on the *g* loading of composites. Due to the increasing presence of brief intelligence tests and abbreviated IQs from full-length intelligence tests, this question is one of great significance. The first step in evaluating the effect of subtest number was to plot the mean *z-g* loading of composites at each level of the subtest variable. See Figure 2. While the mean *z-g* loading of composites increased in a linear fashion from 2 to 7 subtests, a less significant increase appeared between 7 subtests and 10 subtests. Further, the use of 15 subtests appeared to have a negative impact on mean *z-g* loading.

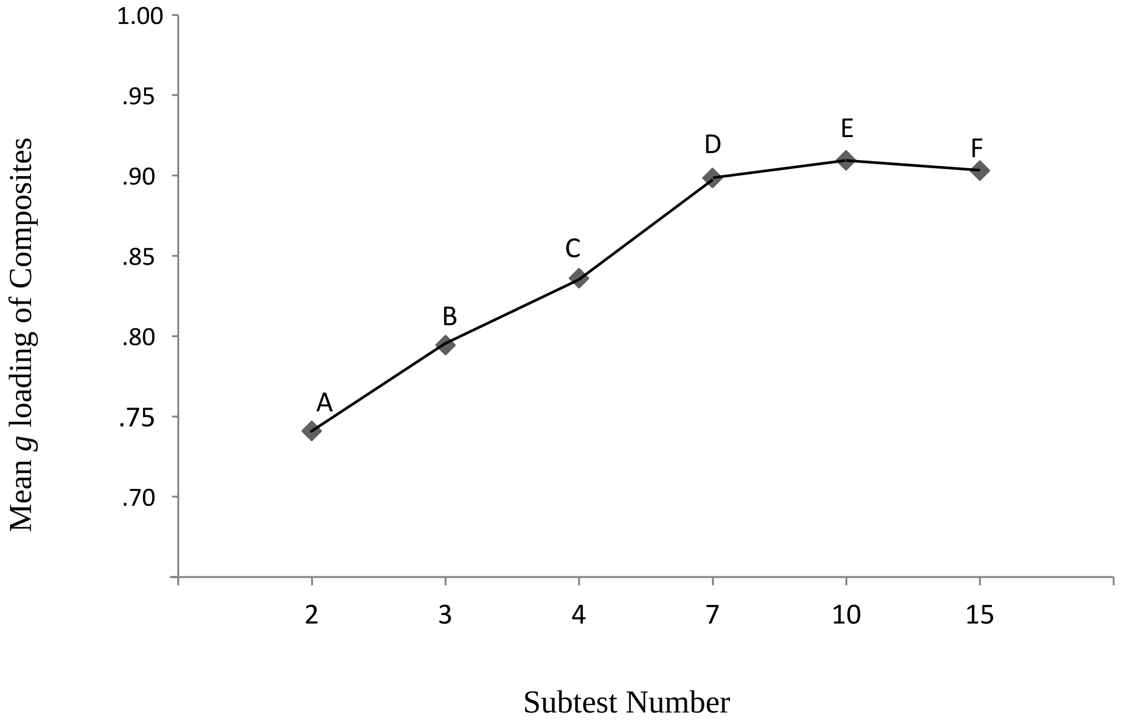

Figure 2. Plot of Mean Fisher z-transformed *g* loading of Composites by Subtest number. Between group difference for A-C, A-D, A-E, A-F, B-D, B-E, B-F, C-D, & C-E were statistically significant at $p < .001$; between group difference for C-F is significant at $p = .001$. Interaction between points A-B, B-C, D-E, D-F are nonsignificant.

The second step in assessing the impact of subtest number on composite *z-g* loading was to determine whether or not graphed differences were statistically significant. A one-way ANOVA was conducted to compare the effect of subtest number, at varying levels, on composite *z-g* loadings. A Tukey correction was made to an alpha level of .05 to account for multiple comparisons. The overall model was statistically significant, $F(5,240) = 42.723$, $p < .001$. Composites created from 2 subtests were significantly lower than composites created from 4, 7, 10, and 15 subtests. A similar pattern appeared for composites created from 3 subtests: they were not significantly different from composites varying by only one subtest, but were significant from composites varying by two or more subtests. Composites created from 4 subtests were significantly higher than composites created from 2 subtests and significantly lower than composites created from 7, 10, and 15 subtests; they were not significantly different from those composites created from 3 subtests. Composites created from 7- 10- or 15-subtests were significantly higher than composites created from 4 or fewer subtests but were not statistically different from one another.

Furthermore, the impact of subtest number was assessed when subtest selection was randomized. Similar to the results presented in Figure 2, the mean *z-g* loading of composites steadily increased from 2 to 7 subtests; a less significant increase appeared between 7 subtests and 10 subtests. In contrast to the data presented in Figure 2, the use of 15 subtests appears to have a positive impact on mean *g* loading.

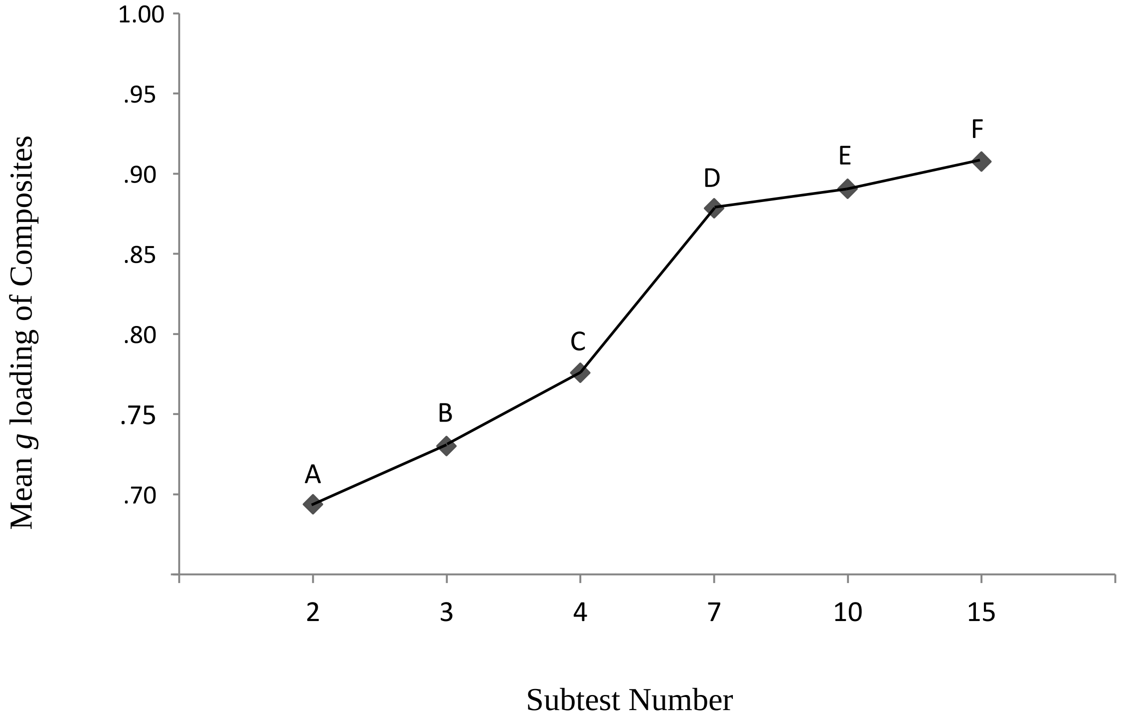

Figure 3. Scatterplot of Mean Fisher z-transformed *g* loading of Randomized Composites by Subtest number. Between group difference for A-D, A-E, A-F, B-D, B-E, and B-F were statistically significant at $p < .001$; between group difference for C-E is significant at $p = .001$; and between group difference for C-D and C-F are significant at $p = .003$. Interaction between points A-B, A-C, B-C, D-E, and D-F are nonsignificant.

The next step in assessing the impact of subtest number on the *g* loadings of randomly generated composites was to determine whether or not graphed differences were significant. An ANOVA was conducted to compare the effect of subtest number, at varying levels, on composite *z-g* loading. As with the data from the full data set, the comparisons on the randomly generated composites resulted in a model that was significant, $F(5,56) = 16.168$, $p < .001$. In contrast to the data from theoretically based composites, randomly generated composites created from 2 subtests were significantly lower than composites created from 7, 10, and 15 subtests, but not from those created from 3 or 4 subtests. A similar pattern appeared for composites created from 3 subtests: they were not significantly different from composites varying by only one subtest but were statistically significant from composites varying by two or more subtests. Composites created from 4 subtests were not significantly different from composites created from 2 or 3 subtests but were significantly lower than composites created from 7, 10, and 15 subtests. Once again, composites created from 7, 10, or 15 subtests were not significantly different from one another.

Heterogeneity

Another goal of this research was to assess the validity of the aggregation hypothesis; the idea that choosing a variety of subtests that target diverse broad abilities would increase the effectiveness of the composite at measuring *g*. The simple test of this hypothesis is that as subtest number increases for composites in which heterogeneous subtest selection was used, the *g* loading of the composite would increase. In the overall model, either when all data were considered or when the data were limited to those composites with 7 or fewer subtests, the impact of the heterogeneity subtest selection method was nonsignificant (see Tables 6 and 7).

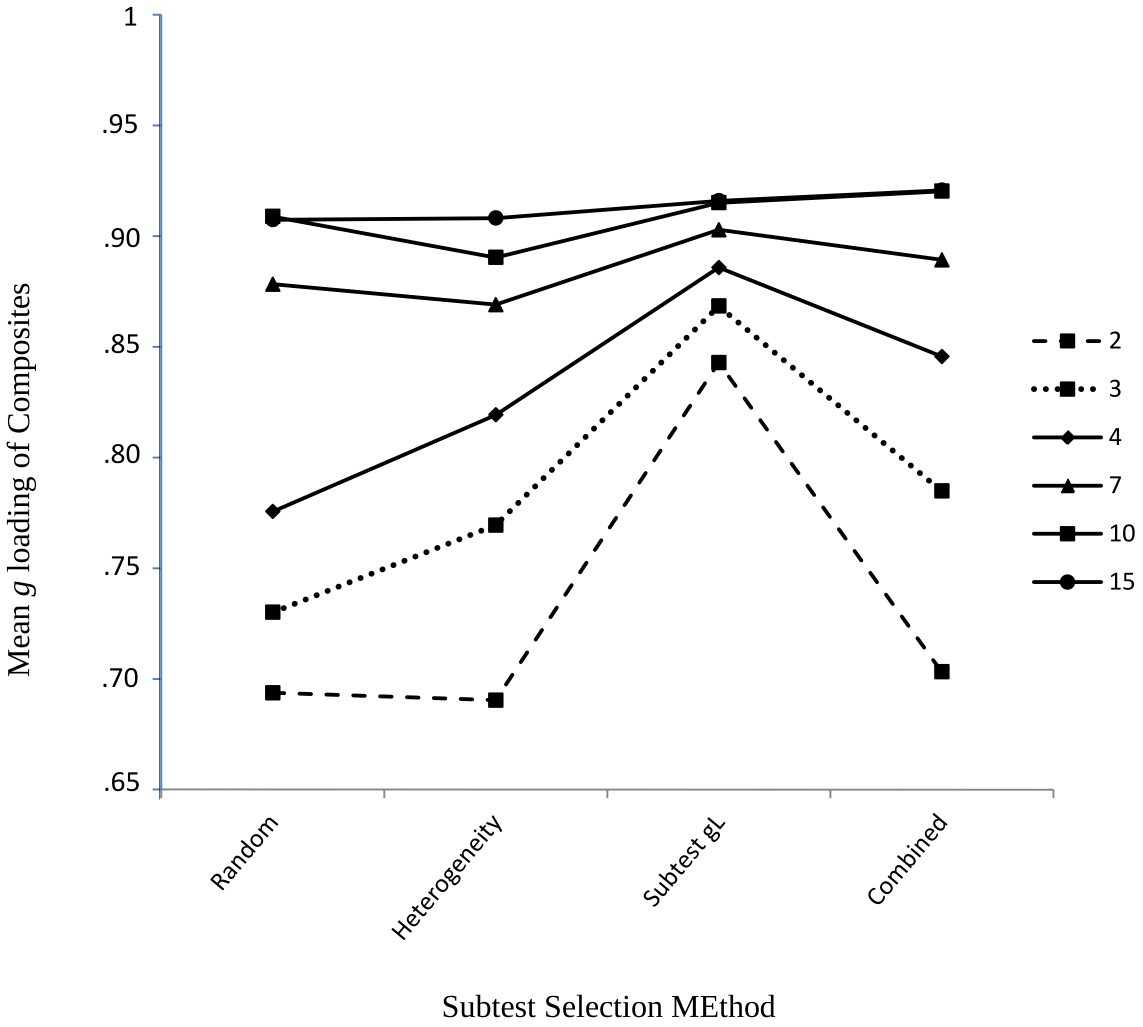

Figure 4. Mean *g* loading by Subtest Selection Method and Subtest Number. Subtest gL = Subtest *g* loading subtest selection method.

Weighting

The fourth hypothesis provided was that subtest weighting would have a positive effect on composite *g* loadings. As indicated by the overall model (See Table 7) in Model 3, weighting had a negligible effect on model prediction.

In order to assess weighting's impact on IQ *g* loadings more directly and to control for subtest selection method, the effect of weighting was evaluated for randomly generated composites. A six (subtest number) by two (weighting) ANOVA was completed, comparing each level of subtest number at each level of weighting. Once again, the simple effect of weighting was nonsignificant, *F*(1,62)=.251, *p* = .619, as was the interaction of weighting and subtest number, *F*(5,62)=.256, *p* = .935.

Aggregation Hypothesis

A final goal was to assess whether or not the interaction between heterogeneity and subtest number significantly impacted IQ *g* loadings. In particular, it was hypothesized that the interaction between these two variables would result in a positive increase in composite *z-g* loading beyond that predicted by either variable independently. To evaluate this hypothesis, an interaction term was created by computing the product of heterogeneity and subtest number. The sequential regression from the overall model using all possible composites—including composites stemming from more than 7 subtests—was used (see Table 6). In a fourth stage, the interaction term between subtest number and heterogeneity was entered by itself and resulted in an nonsignificant increase in model prediction, ΔR^2=.008, *F*(1,238)=3.838, *p* = .051. These additional data are presented in Table 8.

Table 8

Coefficient Values when z-g Loading for Composites from 7 or Fewer Subtests is Sequentially Regressed on Sample, Subtest Number, Subtest Selection Method, Weighting, and the Interaction Between Subtest Number and Heterogeneity.

Model	*B*	SE *B*	β	*B* 95% CI	ΔR^2
1					.020
Sample	.051	.026	.140	[-.001, .102]	
2					$.399^*$
Subtest Number	.100	.009	.632	[.083, .118]	
3					$.189^*$
Heterogeneity	.041	.039	.069	[-.035, .118]	
g Loading	.334	.039	.563	[.258, .410]	
Weighting	-.001	.027	-.002	[-.055, .053]	
Combined	-.231	.055	-.336	[-.338, .123]	
4					.008
IX(Het*Subtest)	.016	.008	.187	[.000, .032]	

Note. *Models 2 and 3 were statistically significant at the *p* < .001 level. Models 1 and 4 were nonsignificant (*p* = .053, and *p* = .051, respectively). Step 3 introduces subtest selection methods.

A six (subtest number) by two (heterogeneity selection method) ANOVA was run in order to fully investigate the relationship between heterogeneity and subtest number. Once again, the simple effect of heterogeneity was nonsignificant, *F*(1,234)=1.730, *p* = 1.90; the interaction between heterogeneity and subtest was also nonsignificant, *F*(5,234)=1.477, *p* = .198.

Discussion

Intelligence tests, regardless of the theory used to develop them, produce IQs that are used in a variety of applied settings for diagnostic purposes, treatment planning, and eligibility determinations. Intelligence tests, through the continued use of IQs, attempt to measure psychometric *g*. Methods used in the creation of IQs have been relatively stable since deviation IQs, scores based on probability estimates, were adopted in the 1960s by test developers following the lead of Wechsler's Bellevue Intelligence Scale (Wechsler, 1939). Methods for creating item content, evaluating items characteristics (i.e., item response theory; for review, see Baker & Kim, 2004), extrapolating normative data to larger populations, and evaluating the potential of cultural bias have indeed progressed significantly, but the only major contributions to methods for creating the IQs themselves have come in bits and pieces and are not wholly adopted. Test developers such as Schrank, McGrew, and Woodcock (2001) began creating IQs by using principal component analysis to develop weighted IQs, while authors such as Elliott (2007) chose a different route, selecting subtests based on their *g* loading.

We must strive for IQs to be as theoretically pure as possible; that is, we must ensure that the measures we use demonstrate significant construct validity (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). The purpose of this study was to evaluate the methods used to create composite scores as a whole and to fill a procedural gap in the science of applied intelligence. Farmer et al. (2013) and Reynolds et al. (2013) showed that IQs are very strong indicators of psychometric *g*. However, Farmer et al (2013) argued that it is "unclear exactly what characteristics lead to higher or lower IQ *g* loadings" (p. 14). Applied intellectual science should strive to better understand what characteristics lead to a more robust relation between IQs and psychometric *g*.

Overall, the results of this study demonstrated the importance of carefully developing composite scores that will be used in interpretation. Several procedures used, either directly or indirectly, by mainstream test developers were not shown to have a significant impact on IQ *g* loadings. For years, the aggregation hypothesis has been the primary theory driving subtest selection for composites; indeed, our research seems to indicate that as the number of subtests used to create an IQ increases, the construct validity of that IQ increases. Essentially, for every subtest added to a composite, the IQ *g* loading was found to increase by .10. However, the effect of subtest number met with diminishing returns beyond seven subtests; composites created from 7, 10, and 15 subtests were not significantly different from one another across all subtest selection methods. Limiting the number of subtests required in order to achieve a construct valid IQ has practical benefits, including reduced client and administrator fatigue and reduced administration times. However, the exact number of subtests that result in maximal relations between IQs and psychometric *g* is unclear, and this question would benefit from additional research. The aggregation hypothesis is two-fold, integrating both the maximization if item and subtest number as well as content.

In contrast to the findings with subtest number, the method of selecting subtests to maximize heterogeneity of the cognitive abilities being sampled had a negligible impact on the construct validity of the IQs when the influence of subtest number was previously controlled. Despite Gustafson's (2002) argument that greater subtest heterogeneity should result in a more precise measure of *g,* our results indicate that heterogeneous subtest selection, per se, had a negligible influence on IQ *g* loadings. Furthermore, our results indicate that heterogeneous subtest selection may result in weakened construct validity when used for abbreviated IQs or IQs from brief intelligence tests, where the number of subtests is too low to allow for expansive cognitive ability sampling. However, the interaction of heterogeneity and subtest number was not robust and may not generalize to other samples. It appears that although the aggregation hypothesis was accurate in predicting stronger construct validity overall, the results of this study seem to indicate that the reasoning for accurate prediction has more to do with the basic tenets of classical true-score theory—that is, more items results in higher reliability—than with the necessity of sampling from a variety of cognitive abilities (e.g., heterogeneity).

Alternatively, the method of selecting subtests to contribute to IQs based on their *g* loading magnitudes clearly results in significantly higher IQ *g* loadings, indicating that it may be the preferred method for ensuring robust IQ–psychometric *g* relations. Furthermore, when an IQ stemmed from few subtests (e.g., 2 or 3), the subtest *g* loading selection method's influence on IQ *g* loadings was even more pronounced, which indicates that this method may be ideal for use with IQs from brief intelligence tests and abbreviated IQs from full-length tests.

The use of weighting methods to produce IQs, as exemplified by the WJ III COG (McGrew & Woodcock, 2001), would ideally produce IQs that account for all of the shared variance in a collection of tests (Schrank et al., 2001) and would produce measures of psychometric *g* that are theoretically pure and devoid of residual influences. The results of this study show, however, that weighting produces IQs that are not more or less *g* loaded than IQs produced from simply summing subtest scores. Thus, these results show that weighting procedures do not enhance the measurement of psychometric *g*, in contrast to our hypothesis.

Combining heterogeneous and subtest *g* loading magnitude selection methods produced a negative impact on IQ *g* loadings. It is not immediately clear why this result may be, but one potential cause is that by first randomly selecting broad ability areas and then choosing the highest *g* loading subtest within those areas, composites were required to feature subtests from areas that are known to be populated by subtests with relatively low *g* loadings (e.g., processing speed).

Previous studies have found IQ *g* loadings for published IQs ranging between .88 and .91 (Farmer et al., 2013 and Reynolds et al., 2013). The present study found that experimental IQs had *g* loadings ranging between .49 and .98. Although comparisons across these studies is less than ideal, the maximum IQ *g* loadings found from published IQs were somewhat lower, but perhaps not statistically or practically different, than the maximum IQ *g* loadings when the characteristics of those IQs were experimentally controlled. This pattern suggests that although we appear to be approaching an upper-limit (nearing perfect ratios of 1.0), there may still be room for improvement in our measurement of psychometric *g* via IQs.

Limitations

A number of limitations are present in this study, including limitations related to (a) independence of data, (b) sampling error, (c) methodology, and (d) the nature of psychometric *g* itself. First, non-independence of observations may be a significant limitation due to drawing *g* loadings from composites constructed from the same cases within samples. As a result, standard errors produced as part of the sequential regression analysis may be biased, and the results of statistical significance testing (e.g., *p* values) may be unreliable. Future research should account for the non-independence of observations through clustering of data at the sample, test, or individual levels.

Second, regarding sampling error and limitations inherent to the archival data used for this project, samples 2 and 3 were somewhat age- and range-restricted and were collected in such a way that the data may be geographically limited, resulting in reduced generalization of our findings. Although we utilized three samples, the modest sample sizes of our data limit our ability to generalize our results and the power of our statistical significance tests. Reeve and Blacksmith (2009) indicate the need for substantial sample size in the modeling and interpretation of latent variables. Reynolds and colleagues (2013) methods were substantially more representative, and future researchers should strive to obtain more generalizable data.

Third, regarding the methodology of our study and of the foundational studies, a number of limitations must be identified. First and foremost, although the general factors modeled in Floyd et al. (2013) were essentially perfectly correlated, two of the samples resulted in correlations that were not perfect (.95 and .97, see Floyd et al., 2013). Due to the fact that *g* to *g* correlations were not perfect across all samples, it is logical that this imprecision may affect the *g* loadings reported in this study. Finally, it has been shown that the IQ–*g* relation decreases as psychometric *g* increases (Reynolds, 2013 and Tucker-Drob, 2009), a well-studied observation known as Spearman's Law of Diminishing Returns (SLODR); however, these effects were not assessed as part of this research, and they may have influenced the *g* loadings reported in this study.

Furthermore, regarding methodology, although subtest number was varied experimentally, it was not done so at every level (i.e., 1 subtest, 2 subtest, etcetera), but rather selected based on common use in mainstream intelligence testing. It is unclear from this research the full nature of the relationship between subtest number and IQ *g* loadings. Finally, a number of test construction methods (e.g., the use of cornerstone items such as those used in the Stanford-Binet Intelligence Scales, Fifth Edition; see

Roid, 2003) were not considered, meaning that this study does not cover the full gamut of IQ construction characteristics.

Finally, as Reynolds and colleagues (2013) also indicated, this study does not attempt to assess for the existence of psychometric *g;* but rather, examines previous research and sets forth to test whether IQ construction methods lead to higher levels of psychometric *g.* As such, the assumption that psychometric *g* is a primary source of variance in cognitive ability measures was made.

Implications

Previous research has shown that although all tests appear to measure the same construct, psychometric *g* (Floyd et al., 2013; Johnson et al., 2008; Keith et al., 2001), the IQs produced by those tests are not perfect indicators of psychometric *g* (Farmer et al., 2013; Reynolds et al., 2013). These results indicate that some methods used in constructing IQs greatly influence how well these IQs represent psychometric *g*. For practitioners using tests of intelligence when an IQ is necessary for diagnosis, selecting tests with IQs that are more likely to have a robust IQ–*g* relation may result in a more accurate measure of an individual’s general intelligence (e.g., intellectual disability assessment).

IQs from full-length intelligence tests tend to be constructed in a variety of different ways with some tests focusing on heterogeneity (e.g., WJ III COG; McGrew & Woodcock, 2001), others focusing on *g* loading (e.g., DAS-2; Elliott, 2007), and still others falling in the middle (e.g., WISC-IV; Wechsler, 2003). Each of these measures, as shown by Farmer et al. (2013) and Reynolds et al. (2013) seem to be adequately measuring psychometric *g*, despite their varied approaches. Certainly as we continue to advance the field of applied intelligence, we will remove those methods not shown to be effective from use. For example, based on the results of this study, weighting appears to be an ineffective procedure and produces no-more valid scores than would simply summing of subtest scores, and it may well be a vestigial characteristic in its current use. Alternatively, emphasizing the *g* loading of subtests seems to have a robust influence on IQ *g* loadings, and it is an area warranting of further research. Furthermore, when a brief IQ or an abbreviated IQ is to be used, these results inform practitioners that selecting tests that focus on ensuring that only the highest *g* loaded subtests are administered may produce more theoretically valid results.

Finally, these findings should be considered by test developers as they develop IQs for intelligence tests. Specifically, when developing brief IQs and abbreviated IQs from full-length intelligence tests, these results indicate that it would be most appropriate to select the subtests that will contribute to an IQ based on their individual subtest *g* loadings. However, in order to maximize the construct validity of IQs, full-length multidimensional tests should aim to draw scores from 7 to 10 subtests. Based on the results of this study, it is unclear, however, whether or not adding additional subtests beyond 10 would have a positive or negative impact on the measurement of psychometric g. Additional research examining the number of subtests—and perhaps even the number of items—contributing to IQs may further inform the development of these global composite scores.

Conclusion

Test developers have developed a variety of methods for calculating global composites scores for intelligence tests. These methods are based either on theory (e.g., the aggregation hypothesis; Rushton et al., 1983) or statistical procedure (e.g., differential weighting of subtest scores; Schrank et al., 2001). The goal of this study was to evaluate the current approaches to IQ formation to determine which of these approaches, if any, resulted in more accurate measurement of psychometric *g*. Study findings indicate that whereas some methods are extremely effective (e.g., increasing subtest number and selecting subtests based on their *g* loadings), others are ineffective or even detrimental (e.g., weighting and heterogeneity). In this study, we have bridged a practical gap in the development of composite scores for intelligence tests.

References

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Long Grove, IL: Waveland Press, Inc.

American Association on Intellectual and Developmental Disabilities. (2010). *Mental retardation: Definition, classification, and systems of supports* (11th ed.). Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (Revised 4th ed.). Washington, DC: Author.

Baker, F. B., & Kim, S. (2004). *Item response theory: Parameter estimations techniques*. (2nd ed.). New York, NY: Marcel Dekker, Inc.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48,* 5–37.

Bickley, P. G., Keith, T. Z., & Wolfe, L. M. (1995). The three-stratum theory of cognitive abilities: Test of the structure of intelligence across the life span. *Intelligence, 20*, 309-328.

Canivez, G. L. (2013). Psychometric versus actuarial interpretation of intelligence and related aptitude batteries. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwean (Eds.), *The oxford handbook of child psychological assessment* (pp. 84-112) Oxford University Press.

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

Chen, H.-Y., Chen, Y.-H., Keith, T. Z., & Chang, B.-S. (2009). What does the WISC-IV measure? Validation of the scoring and CHC-based interpretative approaches. *Journal of Research in Education Sciences, 54*, 85-108.

DiStefano, C., Zhu, M., & Mîndrilă, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation, 14*(20). Available online: http://pareonline.net/getvn.asp?v=14&n=20

Dombrowski, S. C., Watkins, M. W., & Brogan, M. J. (2009). An exploratory investigation of the factor structure of the Reynolds Intellectual Assessment Scales (RIAS). *Journal of Psychoeducational Assessment, 27*, 494-507.

Elliott, C. (1990). *Differential Ability Scales*. San Antonio, TX: Psychological Corporation.

Elliott, C. (2007). *Differential Ability Scales, Second Edition*. San Antonio, TX: Psychological Corporation.

Elliott, C. (2012). The Differential Ability Scales – Second Edition. In Dr. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed.; pp. 336-356). New York: Guilford Press.

Farmer, R. L., Floyd, R. G., Reynolds, M. R., & Kranzler, J. (2013). *The g loadings of IQs across intelligence tests with children and adolescents: IQs are very strong but not perfect indicators of psychometric g.* Unpublished manuscript.

Floyd, R. G., Clark, M. H., & Shadish, W. R. (2008). The exchangeability of IQs: Implications for professional psychology. *Professional Psychology: Research and Practice, 39*, 414–423.

Floyd, R. G., McGrew, K. S., Barry, A., Rafael, F. A., & Rogers, J. (2009). General and specific effects on Cattell–Horn–Carroll broad ability composites: Analysis of the Woodcock–Johnson III Normative Update CHC factor clusters across development. *School Psychology Review, 38*, 249–265.

Floyd, R. G., Reynolds, M. R., Farmer, R. L., & Kranzler, J. H. (2013). Are the general factors from different child and adolescent intelligence tests the same? Results from a five-sample, six-test analysis. *School Psychology Review*.

Gottfredson, L. S. (1998). The general intelligence factor. *Scientific American Presents – Exploring Intelligence, 9*(4), 24-29.

Gottfredson, L. S. (2003). The challenge and promise of cognitive career assessment. *Journal of Career Assessment*, *11*(2), 115-135.

Gustafsson, J. E. (2002). Measurement from a hierarchical point of view. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 73-95). Mahwah, NJ: Erlbaum.

Haahr, M. (2012). Random.org. Retrieved from http://www.random.org.

Homack, S. R., & Reynolds, C. R. (2007). Essentials of assessment with brief intelligence tests. In A. S. Kaufman, & N. L. Kaufman (Series Eds.). *Essentials of psychological assessment.* New York: Wiley.

Humphreys, L. G. (1979). The construct of general intelligence. *Intelligence*, 3, 105–120.

Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C. § 1401. (2004).

Jensen, A. R. (1998). *The g factor*. Westport, CT: Praeger Publisher.

Johnson, W., Bouchard, T. J., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one g factor: Consistent results from three test batteries. *Intelligence, 32*, 95-107.

Johnson, W., Nijenhuis, J. T., & Bouchard, T. J. (2008). Still just 1 g factor: Consistent results from five test batteries. *Intelligence, 36*, 81-95.

Keith, T. Z., Fine, J. G., Reynolds, M. R., Taub, G. E., & Kranzler, J. H. (2006). Higher order, multisample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children—Fourth edition: What does it measure? *School Psychology Review, 35*, 108-127.

Keith, T. Z., Kranzler, J. H., & Flanagan, D. P. (2001). What does the Cognitive Assessment System (CAS) measure? Joint confirmatory factor analysis of the CAS and the Woodcock Johnson Tests of Cognitive Ability (3rd edition). *School Psychology Review, 30,* 89-119.

Keith, T. Z., Low, J. A., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2010). Higher-order factor structure of the Differential Ability Scales—II: Consistency across ages 4 to 17. *Psychology in the Schools, 47,* 676-697.

Keith, T. Z., Quirk, K. J., Schartzer, C., Elliott, C. D. (2009). Construct bias in the Differential Ability Scales? Confirmatory and hierarchical factor structures across three ethnic groups. *Journal of Psychoeducational Assessment, 17,* 249-268.

Kranzler, J., & Floyd, R. G. (2013). *A practical guide to assessing intelligence*. New York: Guilford Press

Maynard, J. L., Floyd, R. G., Acklie, T. J., & Houston L., III. (2011). General factor loadings and specific effects of the Differential Ability Scales, Second Edition composites. *School Psychology Quarterly, 26*, 108-118.

McGrew, K. S., & Flanagan, D. P. (1998). *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment*. Boston: Allyn & Bacon.

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock–Johnson III technical manual.* Itasca, IL: Riverside Publishing.

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Phelps, L., McGrew, K. S., Knopik, S. N., & Ford, L. A. (2005). The general (*g*), broad, and narrow CHC stratum characteristics of the WJ III and WISC-III tests: A confirmatory cross-battery investigation. *School Psychology Quarterly, 20,* 66–88.

Reeve, C. L., & Blacksmith, N. (2009). Equivalency and reliability of vectors of *g*-loadings across different methods of estimation and sample sizes. *Personality and Individual Differences, 47,* 968–972.

Reynolds, C. R., & Kamphaus, R. W. (2003). *Reynolds Intellectual Assessment Scales*. Lutz, FL: Psychological Assessment Resources Inc.

Reynolds, M. R., Floyd, R. G., & Niileksela, C. R. (2013). *How well is psychometric g indexed by global composites? Evidence from three popular intelligence tests.* Unpublished manuscript.

Roid, G. H. (2003). *The Stanford–Binet Intelligence Scales, Fifth Edition*. Itasca, Il: Riverside Publishing.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94*, 18-38.

Sanders, S., McIntosh, D. A., Dunham, M., Rothlisberg, B. A., & Finch, H. (2007). Joint confirmatory factor analysis of the Differential Ability Scales and the Woodcock–Johnson Tests of Cognitive Abilities–Third Edition. *Psychology in the Schools, 44,* 119–138.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*, *86*(1), 162-173.

Schrank, F. A., McGrew, K. S., & Woodcock, R. W. (2001). *Technical abstract* (Woodcock–Johnson III Assessment Service Bulletin No. 2). Itasca, IL: Riverside Publishing.

Sternberg, R. J., & Grigorenko, E. L., & Bundy, D. A. (2001). The predictive value of IQ. *Merrill-Palmer Quarterly*, *47*(1), 1-41.

Taub, G. E., & McGrew, K. S. (2004). A confirmatory factor analysis of Cattell-Horn-Carroll theory and cross-age invariance of the Woodcock–Johnson Tests of Cognitive Abilities III. *School Psychology Quarterly, 19*(1), 72-87. doi: 10.1521/scpq.19.1.72.29409

te Nijenhuis, J., & van der Flier, H. (2005). Immigrant-majority group differences on work-related measures: the case for cognitive complexity. *Personality and Individual Differences, 38,* 1213-1221.

Tucker-Drob, E. M. (2009). Differentiation of cognitive abilities across the life span. *Developmental Psychology*, *45*, 1097-1118. doi:10.10997/a0015864

von Stumm, S., Hell, B., & Chamorro-Premuzic, T. (2011). The "hungry mind": Intellectual curiosity as third pillar of intellectual competence. *Perspectives on Psychological Science, 6,* 574–588.

Watkins, M. W., Glutting, J. J., & Lei, P.-W. (2007). Validity of the full-scale IQ when there is significant variability among WISC-III and WISC-IV factor scores. *Applied Neuropsychology*, *14*, 13-20.

Wechsler, D. (1939). *The measurement of adult intelligence*. Baltimore: Williams & Wilkins.

Wechsler, D. (1991). *The Wechsler Intelligence Scale for Children, Third Edition.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003). *The Wechsler Intelligence Scale for Children–Fourth Edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2011). *The Wechsler Abbreviated Scale of Intelligence, Second Edition*. San Antonio, TX: Psychological Corporation.

World Health Organization, Division of Mental Health And Prevention of Substance Abuse. (2010). *ICD-10 guide for mental retardation.* Geneva: World Health Organization.

Ryan Farmer <r.farmer27@gmail.com>

RE: Request to extend a data agreement DAS-II, WISC-IV, and KABC-II Amendment

Randy G Floyd (rgfloyd) <rgfloyd@memphis.edu> Mon, Dec 12, 2011 at 6:17 PM
To: "HAS-SAT Shared Dist. and Licensing" <pas.Licensing@pearson.com>
Cc: "r.farmer27@gmail.com" <r.farmer27@gmail.com>

Bill, as usual, I appreciate your support. Thanks, too for extending the agreement to two years!

With appreciation,

Randy

From: HAS-SAT Shared Dist. and Licensing [pas.Licensing@pearson.com]
Sent: Saturday, December 10, 2011 1:20 PM
To: Randy G Floyd (rgfloyd)
Subject: RE: Request to extend a data agreement DAS-II, WISC-IV, and KABC-II Amendment

Dear Dr. Floyd,

Thank you for your request to reinstate and extend your Standardization Agreement dated January 5, 2009, as amended on December 22, 2009 (collectively "the Agreement") between NCS Pearson, Inc. and yourself for permission to use standardization data from the Differential Ability Scales®−Second Edition (DAS®−II) and the Wechsler Intelligence Scale for Children®− Fourth Edition (WISC®−IV) in your continuing research regarding several research studies: An analysis of g loadings and specificity of the CHC factor cluster scores in the DAS-II; secondly, use DAS-II data to examine clusters (versus factor indexes) and to investigate the cognitive ability profiles of children with mental retardation (MR) using the Wechsler factor indexes; and thirdly, to examine the g saturation of the Wechsler General Ability Indexes (GAI) and Full Scale IQs (FSIQ) using five samples of children and adults who took one version of the Wechsler batteries and another comprehensive cognitive assessment battery in a counterbalanced order. This Second Amendment is effective as of January 6, 2011.

By this email, we are reinstating and amending your Agreement to add an additional twenty-four (24) months.

The Agreement is amended as follows:

Section 1a. of the Agreement is deleted in its entirety and replaced by the following new Section 1a.

1. a. The License granted herein shall be for a period commencing with the February 1, 2007 expiration date of the previous agreement to use the same data for the same purpose, and expiring January 5, 2014, whereupon the Licensed Use shall cease.

All other terms and conditions remain in full force and effect.

Please keep a copy of this email with your Agreement.

Regards,

William H. Schryver
Senior Licensing Specialist
Clinical Assessment
Pearson
19500 Bulverde Road
San Antonio, TX 78259
T: (210) 339-5345
F: (210) 339-5059
E: pas.licensing@pearson.com<mailto:pas.licensing@pearson.com>

Pearson
Always Learning
Learn more at www.psychorp.com

____
From: Randy G Floyd (rgfloyd) [mailto:rgfloyd@memphis.edu]
Sent: Mon 11/28/2011 9:22 PM
To: HAS-SAT Shared Dist. and Licensing
Subject: Request to extend a data agreement DAS-II, WISC-IV, and KABC-II Amendment

Bill,

I hope that you are doing well.

I am writing to request another extension on a data agreements you have previously granted. We have completed one of the studies for which we requested data, one article is still under review in a peer-reviewed journal, and other projects are in their final stages of completion. (Please let me know if you need more specifics.)

I would like to request that the data agreement described below be extended; it focused on the DAS-II and WISC-IV. A second one--granting permission to use data we originally obtained from AGS--for the KABC-II was granted on October 25, 2010.

I am willing to provide additional information to you as needed.

As always, thanks for your assistance,

Randy

Randy G. Floyd, Ph.D.
Editor, Journal of School Psychology
Associate Professor of Psychology
The University of Memphis
202 Psychology Building
400 Innovation Drive
Memphis, TN 38152
Phone: 901-678-4846
Fax: 901-678-2579

From: HAS-SAT Shared Dist. and Licensing [pas.Licensing@pearson.com]
Sent: Tuesday, December 22, 2009 9:37 AM
To: Randy G Floyd (rgfloyd)
Subject: Request to extend a data agreement DAS-II & WISC-IV that expires January 5, 2010 - 1st Amendment

Dear Dr. Floyd,

Thank you for your request to reinstate and extend your Standardization Agreement dated January 5, 2009 ("the Agreement") between NCS Pearson, Inc. and yourself for permission to use standardization data from the Differential Ability Scales®−Second Edition (DAS®−II) and the Wechsler Intelligence Scale for Children®− Fourth Edition (WISC®−IV) in your continuing research regarding three research studies: An analysis of g loadings and specificity of the CHC factor cluster scores in the DAS-II; secondly, use DAS-II data to examine clusters (versus factor indexes) and to investigate the cognitive ability profiles of children with mental retardation (MR) using the Wechsler factor indexes; and thirdly, to examine the g saturation of the Wechsler General Ability Indexes (GAI) and Full Scale IQs (FSIQ) using five samples of children and adults who took one version of the Wechsler batteries and another comprehensive cognitive assessment battery in a counterbalanced order.

By this email, we are amending your Agreement to add an additional twelve (12) months.

The Agreement is amended as follows:

Section 1a. of the Agreement is deleted in its entirety and replaced by the following new Section 1a.

1. a. The License granted herein shall be for a period commencing with the February 1, 2007 expiration date of the previous agreement to use the same data for the same purpose, and expiring January 5, 2011, whereupon the Licensed Use shall cease.

All other terms and conditions remain in full force and effect.

Please keep a copy of this email with your Agreement.

Regards,

Bill Schryver
William (Bill) Schryver
Permissions Specialist
Clinical Assessment
Pearson
19500 Bulverde Rd
San Antonio, TX 78259-3701
Tel. 210-339-5345 or 800-228-0752 ext 5345

Fax. 210-339-5601
pas.licensing@pearson.com<mailto:pas.lpas.licensing@pearson.com>

From: Randy G Floyd (rgfloyd) [mailto:rgfloyd@memphis.edu]
Sent: Mon 12/21/2009 3:35 PM
To: HAS-SAT Shared Dist. and Licensing
Subject: RE: Request to reinstate and extend a data agreement WAIS-III & WISC-IV that expired October 31, 2009 - 1st Amendment

Bill, thanks for the amazingly quick response to my request. I appreciate it.

I also have another such agreement with a January 5, 2010 end date. Could I please have an extension on that one as well? (I apologize for not requesting it along with the other. It was in a different file.)

Happy holidays to you!

Thank you,
Randy

From: HAS-SAT Shared Dist. and Licensing [pas.Licensing@pearson.com]
Sent: Monday, December 21, 2009 9:47 AM
To: Randy G Floyd (rgfloyd)
Subject: Request to reinstate and extend a data agreement WAIS-III & WISC-IV that expired October 31, 2009 - 1st Amendment

Dear Dr. Floyd,

Thank you for your request to reinstate and extend your Standardization Agreement dated October 28, 2008 ("the Agreement") between NCS Pearson, Inc. and yourself for permission to use standardization data from the Wechsler Adult Intelligence Scale®−Third Edition (WAIS®−III) and the Wechsler Intelligence Scale for Children® − Fourth Edition (WISC®−IV) in your continuing research regarding the cognitive ability profiles of children with mental retardation.

By this email, we are amending your Agreement to reinstate it and to add an additional twelve (12) months.

The Agreement is amended as follows:

Section 1a. of the Agreement is deleted in its entirety and replaced by the following new Section 1a.

1a. The License granted herein shall be for a period commencing with the February 1, 2007 expiration date of the previous agreement to use the same data for the same purpose, and expiring October 31, 2010, whereupon the Licensed Use shall cease.

All other terms and conditions remain in full force and effect.

Please keep a copy of this email with your Agreement.

Regards,

Bill Schryver
William (Bill) Schryver
Permissions Specialist
Clinical Assessment
Pearson

19500 Bulverde Rd
San Antonio, TX 78259-3701
Tel. 210-339-5345 or 800-228-0752 ext 5345
Fax. 210-339-5601
pas.licensing@pearson.com<mailto:pas.lpas.licensing@pearson.com>

From: Randy G Floyd (rgfloyd) [mailto:rgfloyd@memphis.edu]
Sent: Sun 12/20/2009 11:03 PM
To: HAS-SAT Shared Dist. and Licensing
Subject: Renewal of Data Agreement

Hi, Bill. I hope that the holiday season is treating you well.

I would like to renew my data agreements with you. We are still working on the projects for which we have approval to use standardization and validity study data. It appears as if our current agreements expired on October 28.

When you have time, would you be willing to assist me in renewing these agreements, please?

Thank you for considering my request,

Randy

Randy G. Floyd , Ph.D.
Associate Professor of Psychology
The University of Memphis
202 Psychology Building
400 Innovation Drive
Memphis, TN 38152
Phone: 901-678-4846
Fax: 901-678-2579

winmail.dat
31K

Ryan Farmer <r.farmer27@gmail.com>

FW: Update from K. McGrew at Woodcock-Munoz Foundation

Randy G Floyd (rgfloyd) <rgfloyd@memphis.edu> Thu, Dec 15, 2011 at 3:57 PM
To: "r.farmer27@gmail.com" <r.farmer27@gmail.com>

Ryan, here is the final approval; it's for the WJ III data sets.

Randy

From: Kevin McGrew [iapsych@me.com]
Sent: Thursday, December 15, 2011 3:06 PM
To: Randy G Floyd (rgfloyd)
Cc: iap@earthlink.net
Subject: Re: Update from K. McGrew at Woodcock-Munoz Foundation

Randy. I have reviewed your proposed analysis of archived data provided to you previously by WMF and am pleased to approve this use of the data in question.

I look forward to seeing your results.

Kevin McGrew

Sent from Kevin McGrew's iPad
Kevin McGrew, PhD
Educational Psychologist
Research Director
Woodcock-Munoz Foundation

---

winmail.dat
5K

From: Kevin McGrew [iapsych@me.com]
Sent: Thursday, December 15, 2011 3:06 PM
To: Randy G Floyd (rgfloyd)
Cc: iap@earthlink.net
Subject: Re: Update from K. McGrew at Woodcock-Munoz Foundation

Randy. I have reviewed your proposed analysis of archived data provided to you previously by WMF and am pleased to approve this use of the data in question.

I look forward to seeing your results.

Kevin McGrew

Sent from Kevin McGrew's iPad
Kevin McGrew, PhD
Educational Psychologist
Research Director
Woodcock-Munoz Foundation

Ryan Farmer <r.farmer27@gmail.com>

IRB Approval2563

Institutional Review Board <irb@memphis.edu> Fri, Feb 8, 2013 at 11:06 AM
To: "Ryan Farmer (rlfarmer)" <rlfarmer@memphis.edu>
Cc: "Randy G Floyd (rgfloyd)" <rgfloyd@memphis.edu>

Hello,

The University of Memphis Institutional Review Board, FWA00006815, has reviewed and approved your submission in accordance with all applicable statuses and regulations as well as ethical principles.

PI NAME: Ryan Farmer
CO-PI: John Kranzler, Ph.D. & Matt Reynolds, Ph.D.
PROJECT TITLE:
FACULTY ADVISOR NAME (if applicable): Randy Floyd

IRB ID: #2563
APPROVAL DATE: 2/8/2013
EXPIRATION DATE:
LEVEL OF REVIEW: Exempt

Please Note: Modifications do not extend the expiration of the original approval

Approval of this project is given with the following obligations:

1. If this IRB approval has an expiration date, an approved renewal must be in effect to continue the project prior to that date. If approval is not obtained, the human consent form(s) and recruiting material(s) are no longer valid and any research activities involving human subjects must stop.

2. When the project is finished or terminated, a completion form must be completed and sent to the board.

3. No change may be made in the approved protocol without prior board approval, whether the approved protocol was reviewed at the Exempt, Exedited or Full Board level.

4. Exempt approval are considered to have no expiration date and no further review is necessary unless the protocol needs modification.

Thank you,

Ronnie Priest, PhD

Institutional Review Board Chair

The University of Memphis.

Note: Review outcomes will be communicated to the email address on file. This email should be considered an official communication from the UM IRB. Consent Forms are no longer being stamped as well. Please contact the IRB at IRB@memphis.edu if a letter on IRB letterhead is required.