Behavior Research Methods (2019) 51:507–522
https://doi.org/10.3758/s13428-018-1098-4

General fluid/inductive reasoning battery for a high-ability population

Patrick Kyllonen[1] · Robert Hartman[2] · Amber Sprenger[2] · Jonathan Weeks[1] · Maria Bertling[1] · Kevin McGrew[3] · Sarah Kriz[2] · Jonas Bertling[1] · James Fife[1] · Lazar Stankov[4]

Published online: 26 November 2018
© Psychonomic Society, Inc. 2018

Abstract

The validity of studies investigating interventions to enhance fluid intelligence (Gf) depends on the adequacy of the Gf measures administered. Such studies have yielded mixed results, with a suggestion that Gf measurement issues may be partly responsible. The purpose of this study was to develop a Gf test battery comprising tests meeting the following criteria: (a) strong construct validity evidence, based on prior research; (b) reliable and sensitive to change; (c) varying in item types and content; (d) producing parallel tests, so that pretest–posttest comparisons could be made; (e) appropriate time limits; (f) unidimensional, to facilitate interpretation; and (g) appropriate in difficulty for a high-ability population, to detect change. A battery comprising letter, number, and figure series and figural matrix item types was developed and evaluated in three large-*N* studies ($N = 3{,}067$, 2,511, and 801, respectively). Items were generated algorithmically on the basis of proven item models from the literature, to achieve high reliability at the targeted difficulty levels. An item response theory approach was used to calibrate the items in the first two studies and to establish conditional reliability targets for the tests and the battery. On the basis of those calibrations, fixed parallel forms were assembled for the third study, using linear programming methods. Analyses showed that the tests and test battery achieved the proposed criteria. We suggest that the battery as constructed is a promising tool for measuring the effectiveness of cognitive enhancement interventions, and that its algorithmic item construction enables tailoring the battery to different difficulty targets, for even wider applications.

Keywords Intelligence · Fluid ability · Gf · Working memory training · Reasoning · Item-response theory · Test assembly

General fluid ability (Gf) is "at the core of what is normally meant by intelligence" (Carroll, 1993, p. 196), and has been shown empirically to be synonymous with general cognitive ability (g), at least within groups with roughly comparable opportunities to learn (Valentin Kvist & Gustafsson, 2008). Gf has been viewed as an essential determinant of one's ability to solve a wide range of novel real-world problems (Schneider & McGrew, 2012). Perhaps because of its association with diverse outcomes, there has been a longstanding interest in improving Gf (i.e., intelligence) through general schooling (Cliffordson & Gustafsson, 2008), direct training (Stankov, 1986), diet (Korol, 2002), cognition-enhancing drugs (Dinges & Weaver, 2003; Ramos & Arnsten, 2007), and various other means (e.g., Kyllonen, Roberts, & Stankov, 2008; Simons et al., 2016). The failure of some earlier efforts to improve intelligence significantly (e.g., Venezuela's Project Intelligence; Herrnstein, Nickerson, de Sánchez, & Swets, 1986) led to a languishing of enthusiasm for improving Gf, despite occasional successes (e.g., Stankov, 1986). But a study by Jaeggi, Buschkuehl, Jonides, and Perrig (2008) reenergized the debate over the improvability of fluid intelligence.

The promise of Gf enhancement has drawn considerable attention in the scientific community, as is shown by the many attempted replications of Jaeggi et al.'s (2008) working memory training (Au et al., 2015; Melby-Lervåg, Redick, & Hulme, 2016). Additionally, non-working-memory interventions, such as novel problem solving (e.g., Stine-Morrow, Parisi, & Morrow, 2008; Tranter & Koutstaal, 2008), video game training (e.g., Basak, Boot, Voss, & Kramer, 2008), and neurostimulation (e.g., Sellers et al., 2015), have been examined (Simons et al., 2016).

✉ Patrick Kyllonen
pkyllonen@ets.org

[1] Educational Testing Service, Princeton, NJ, USA

[2] MITRE Corporation, Bedford, MA, USA

[3] Institute for Applied Psychometrics, St. Joseph, MN, USA

[4] Institute for Positive Psychology & Education, Australian Catholic University, Sydney, New South Wales, Australia

The evidence for the efficacy of Gf-enhancing interventions has been mixed. Several studies have replicated Jaeggi et al.'s (2008) findings (e.g., Jaeggi, Buschkuehl, Shah, & Jonides, 2014; Jaeggi et al., 2010; Jaušovec & Jaušovec, 2012; Karbach & Kray, 2009), but others have not (e.g., Chein & Morrison, 2010; Harrison et al., 2013; Redick et al., 2013). Some meta-analyses have shown that working memory training has small but significant effects on Gf (Au et al., 2015; Karbach & Verhaeghen, 2014; Schwaighofer, Fischer, & Bühner, 2015), but others have concluded that evidence for training effects can be traced to pretreatment differences, the use of passive (vs. active) control groups, and other design weaknesses (Dougherty et al., 2016; Melby-Lervåg et al., 2016; Shipstead, Redick, & Engle, 2012).

A recent review (Simons et al., 2016) provided several recommendations for evaluating the effectiveness of interventions designed to improve Gf. Among these were the use of reliable, multiple measures, pretesting to ensure comparability between the treated and control groups, and the computing of the posttest–pretest differences for the treated versus control groups. We concur, and here we note several features of a Gf measure that are desirable in a study designed to evaluate the effects of a treatment to enhance Gf. These are summarized below.

Tests should be valid measures of Gf

The interpretation of findings from Gf enhancement research depends on how Gf is measured, both prior to and following an intervention. Carroll's (1993) reanalysis of factor analytic studies of cognitive abilities identified Gf as one of eight second-stratum factors and identified three main types of tasks within the reasoning domain: deductive, inductive, and quantitative reasoning, with inductive reasoning having the highest average correlations with the other two categories, suggesting its centrality (Schneider & McGrew, 2012; Wilhelm, 2005). Carroll (1993, Table 6.1, pp. 215–216) showed that matrices and series tasks were the most studied tasks of all reasoning tasks. In this study, we focused on matrices and series tasks.

Tests should be reliable and sensitive to change

It is useful to have reliable Gf scores for evaluating intervention effects. Reliability limits validity; a test with low reliability is a poor indicator of the target construct (Gf). Low reliability reduces the possibility of examining potential moderators of change in pre–post designs. Although unreliable tests can be used to measure change (e.g., effect size estimates are not related to reliability) the interpretation of that change is ambiguous if the instrument used to measure change is unreliable. Consequently, sensitivity-to-change measures proposed in the clinical literature to enable interpretable change are directly related to reliability (Eisen, Ranganathan, Seal, & Spiro, 2007; Jacobson & Truax, 1991). In this study, we targeted $r_{xx'}$ = .90 as our reliability target for the various Gf measures, in line with the reliability of commercial instruments used in high-stakes decision making (e.g., Educational Testing Service [ETS], 2016).

There should be multiple measures, with varied content

A common strategy for Gf-focused enhancement studies has been to use only a single measure to assess Gf, such as Raven's Progressive Matrices (Basak et al., 2008; Chein & Morrison, 2010; Duthie et al., 2002; Hayes, Petrov, & Sederberg, 2015; Jaeggi et al., 2008; Jaeggi et al., 2010). The limits of any single test as an indicator of a latent factor have long been recognized, prompting a call in Gf enhancement studies for expanding Gf measurement to different reasoning problem types and stimulus domains (Buschkuehl & Jaeggi, 2010; Hayes et al., 2015; Jaeggi et al., 2014; Jaeggi et al., 2010; Schneider & McGrew, 2012; Shipstead et al., 2012; Sternberg, 2008; von Bastian & Oberauer, 2013). Here we developed series and matrices items based on verbal, quantitative, and spatial content, requiring only simple transformation rules.

There should be parallel tests for pretest and posttest administration

In repeated testing, it is desirable to administer parallel test forms to ensure that the measurement instruments and their resulting scores are comparable and scores can be placed on a common scale (Kolen & Brennan, 1995). Parallel (in contrast to identical) forms would prevent item-learning effects from confounding Gf enhancement findings.

Tests should have appropriate time limits

Many Gf enhancement studies have downplayed the importance of time limits, which can affect number and percentage-correct scores (Moody, 2009; Redick et al., 2013), affect the construct being measured (Gs; McGrew, 2009), and threaten the validity of the test as a measure of Gf (e.g., Basak et al., 2008; Harrison et al., 2013; Jaeggi et al., 2010; Rae, Digney, McEwan, & Bates, 2003; Redick et al., 2013; Stephenson & Halpern, 2013). Here we provided relaxed time limits to ensure a power test, based on the speed–power literature (Bridgeman et al. 2003; Kyllonen & Zu, 2016).

Individual tests should be unidimensional

Unidimensionality in item responses is an important quality for a test designed to measure a construct, for several reasons. Violations of unidimensionality can (a) render score interpretations ambiguous, (b) reduce internal consistency estimates of reliability, (c) make it more difficult to develop parallel forms, or equate forms, (d) lessen comparability between examinees presenting similar scores as they may achieve them by solving items reflecting different dimensions, and (e) complicate the use of item response theory modeling, which assumes unidimensionality. However, violations of unidimensionality are commonplace. Essential unidimensionality (Stout, 1987) is a less restrictive form in which the weight of the dimensions (in a total score composite) is consistent for scores across the range of abilities tested. The important issue is to avoid large, systematic deviations from unidimensionality that would jeopardize score interpretation and detract from our ability to equate or create parallel forms.

Tests should be difficult enough to avoid ceiling effects

Ceiling effects are especially problematic in pre–post intervention designs, in which the aim is to identify significant increases in cognitive ability. A goal for this study was to produce a test with an adequate measurement range that would allow for sensitivity to changes within individuals (Embretson, 1991), particularly for a high-ability adult population. We did this for several reasons: (a) our targeted population was high-ability; (b) we wished to avoid posttest ceiling effects; (c) our item modeling strategy enabled the creation of easier items through the elimination of processing steps or the reduction of memory burden, so an easier test would be a special case of the test we developed; (d) in the test administered, we randomized item order due to the design, but in practice the test could be made easier by ordering the items from easy to difficult; and (e) because the test administration was in a low-stakes setting, we could expect score increases of approximately 0.5 standard deviations or more with even modest performance incentives (Liu et al. 2012). For this study, we sought to provide reasonably precise measurement at a range of ability up to 3.5 standard deviations above the mean of a population of American third and fourth-year undergraduate students, using highly educated samples.

Purpose of the study

We aimed to develop a Gf test battery that met the aforementioned requirements for detecting Gf change. We used item response theory methods to create a test battery that incorporated a variety of tasks (e.g., series, matrices) and used item types from the three major Gf content domains (verbal, numerical, and spatial). Parallel forms were also developed using test-equating methods (Kolen & Brennan, 1995), in order to enable the attribution of postintervention gains to the treatment rather than to form difficulty differences.

All items, instructions, scoring keys, item statistics, and other supplementary information for all items and tests described in this article are available online (MITRE/ETS, 2016). Additional information can also be obtained in a supplementary technical paper (Weeks, Kyllonen, Bertling, & Bertling, 2017).

Study 1: Series battery development (number, letter, and arrow)

We developed a battery of fluid ability tests based on series reasoning, for several reasons. First, series measures are excellent measures of Gf, among the best that have been evaluated (Carroll, 1993). Series completion tests have been included in many cognitive abilities batteries (Jäger et al., 2006; Thorndike & Hagen, 1971; Thurstone & Thurstone, 1941), including in the ETS Kit of Factor-Referenced Cognitive Tests (Ekstrom, French, Harman, & Dermen, 1976), and as an item type (called *pattern identification*) of the Analytic section of ETS's Graduate Record Examination (GRE) (Bridgeman & Rock, 1993; Emmerich, Enright, Rock, & Tucker, 1991).

Second, series tests can be constructed with different content (verbal, numerical, and spatial) using overlearned stimulus elements, such as letters of the alphabet and low value integers whose relationships are also overlearned, such as identity, successor, and predecessor. The effects of differential education and culture are likely minimized due to the use of simple rules applied to an overlearned stimulus set assuming that participants are highly familiar with the Roman alphabet and with numbers. Third, an extensive information-processing literature on series tests provides automatic item-generation models.

Series item models

Simon and Kotovsky (1963; Kotovsky & Simon, 1973) proposed an initial framework for letter series, and Holzman, Pellegrino, and Glaser (1983) proposed a similar one for number series. Consider the series item, 2 4 6 7 9 11 12 14 __. The components of the framework are *relations detection* (determining the relationship between contiguous elements; here + 2 or + 1), *discovery of periodicity* (finding the length of the period within the longer series; here 3), *completion of the pattern description within the period* (identifying relations between elements within a period; here + 2), and *extrapolation* (use the pattern description to fill in the blank; here, because the blank is the third position of a period, + 2 gives 16).

Simon and Kotovsky (1963) developed a pattern description language (PDL) to define the information processing requirements for solving series problems. Holzman et al. (1983) examined the importance of variables derived from the PDL on problem difficulty (e.g., the length of the item description, period length), and found that working memory placekeepers (WMP; i.e., the number of occasions during a processing sequence on which a number or relationship had to be remembered) accounted for most of the variance in item difficulty (Myors, Stankov & Oliphant, 1989; Stankov & Cregan, 1993; Stankov & Myors, 1990).

Overview of Study 1

We aimed to produce two parallel forms of series items, each consisting of a letter (L), number (N), and arrow (A) series test, as well as a composite (C) test with items from all three content tests. Each test was to be of a length sufficient for a reliability of ≥ .90. On the basis of prior research, we estimated that this would require approximately 30–40 items on each test.

We administered subsets of items to participants in order to estimate item difficulty (b) and item discrimination (a) using the two-parameter logistic (2PL) item response theory model (Birnbaum, 1968).[1] We created item subsets using a balanced incomplete block design similar to those used in large-scale assessments, then used missing data imputation methods to estimate item parameters.

From this calibrated item pool, our goal was to assemble two parallel forms using automated test assembly (ATA) methods (van der Linden, 2005). The main issues we addressed in Study 1 were (a) whether we could generate Gf series items from information-processing descriptions that would satisfy the criteria of providing reliable measurement across a wide ability range; and (b) whether we could find evidence that such measures were valid indicators of Gf as shown by correlations with demographics and self-reported SAT and ACT scores.

Method

Participants

We contracted Qualtrics Labs, Inc., a supplier of participant panels for survey research, to provide 2,000 participants (at $24.50/each), and host the online assessment on their servers. Participants were targeted to have the following educational statuses (400 for each): (a) third- or fourth-year undergraduates, (b) bachelor’s degree holders without a master’s degree, (c) master’s degree students, (d) master’s degree holders without a doctorate, and (e) doctorate degree holders. The participants had to be U.S. citizens, and psychology degree holders were excluded. Educational status levels were distributed evenly across the survey forms.

For Study 1, we also conducted a supplemental study, in which we recruited an additional 1,036 examinees from Qualtrics Labs, Inc., with demographic characteristics similar to those from the main study. A subset of these examinees (n = 157) were respondents who participated in Study 1 but took forms that included arrow and letter series items only.

Measures

We wrote approximately 100 items for each of the three content domains—letter (L), number (N), and arrow (A)—varying in expected difficulty. We selected a subset of 80 items for administration after item reviews by the authors (different authors reviewed different items based on experience and expertise) and by cognitive interviews with research assistants (i.e., we showed items to research assistants and asked them to think aloud to ensure proper understanding of the task). (We developed an additional 50 number series items.) All of the tasks in Study 1 and in all studies were programmed for the Qualtrics Research Suite survey platform.

Letter series (LS) LS items were created by varying (a) series length (7 to 16), (b) period length (2, 3, 4), (c) period position of the blank (1, 2, 3, 4), (d) operators (repeat, + 2), (e) WMP level (1, 2, 3, or 4), (f) letter starting values (A through Z), and (g) the requirement (yes, no) to recycle through the alphabet due to successor relationships on Y and Z (i.e., Y + 2 = A). Examinees were asked to type in the missing letter (a to z). Combinations of design factors (nonexhaustive) created 16 pattern description language (PDL) patterns (i.e., item models), which were sampled from to form one 16-item mini-test. Each mini-test comprised the same 16 PDL patterns,

[1] Because some of our items were multiple-choice (figural matrices, and letter and arrow series offering limited choices) and therefore allowed guessing, reviewers asked about the use of the three-parameter logistic model (3PL), an item response theory model that includes a guessing parameter. However, the identifiability of the 3PL is considered an “open problem” (San Martín, González, & Tuerlinckx, 2015, p. 466). Model identification problems result in problems with consistent and unbiased parameter estimation and parameter interpretation. For this reason, the 2PL is often preferred in operational testing, such as in the OECD’s (2017) Program for International Student Assessment (PISA) and Educational Testing Service’s Graduate Record Examination (Robin, Steffen, & Liang, 2014).

but across mini-tests, the same pattern created a different item because same-PDL-pattern items varied in starting letters, series length, blank period position, and recycling requirement. Five LS mini-tests were developed (L1 to L5).

Arrow series (AS) AS items were analogous to LS items except that the content, instead of the alphabet, was a set of arrows, pointing in the eight cardinal directions (north, northeast, east, southeast, south, southwest, west, northwest). Examinees were asked to type in the missing arrow (using a key indicating a mapping of cardinal directions to the numbers 1 to 8). Successor relationships were defined as 45-deg rotations of arrow positions; thus north + 3 = southeast. Similarly to the number series (NS), AS items were created by varying (a) series length (six to 14), (b) period length (2, 3, 4), (c) period position of the blank (1, 2, 3, 4), (d) operators (repeat, + 2, + 3), (e) WMP level (1, 2, 3, or 4), and (f) starting values (north, northeast, . . . , northwest). As with NS items, there were 16 PDL patterns (each generating seven to 16 items), and items were drawn from the patterns to create five 16-item mini-tests (N1 to N5).

Number series (NS) NS items were generated on the basis of the item models from Bertling (2012) (although the models we used were modified from Bertling's). The period length was always 1, and series consisted of eight elements (integers) with the 9th missing. Examinees were asked to type in the missing integer. Series were generated by applying one or more operators to the current element, in a particular order. Operators were (a) adding or subtracting a constant (+ 1, – 1, + 2, – 2); (b) adding or subtracting a checksum1 (sum of the digits of the element; e.g., 11 => 2); (c) adding or subtracting a checksum2 (sum of the digits across the current and previous elements in the series; e.g., 29, 39 => 23); (d) adding the current with the previous element, as in a Fibonacci series; and (e) treating either the first element alone or the first two elements as the starting values from which to compute the next series element. For example, the series 6, 4, 1, 5, 6, _ could be generated by treating the first two elements as the starting values, then applying a combination of the Fibonacci and checksum1 rules (1 is the checksum1 of 10 [= 6 + 4], 5 is the sum of 4 + 1, 6 is the sum of 1 + 5, and the sixth element would be the checksum1 of 6 + 5, which would be 2 [i.e., the checksum1 of 11]). Each mini-test consisted of 16 item models from varying the design factors, and we created five instances per item model (instances varied on the constant and the one or two starting elements). For the supplemental study, additional NS items were generated from some of the easier item design combinations (due to the first batch being too difficult). These were assembled into three 20-item mini-tests.

Design

Items were assembled into same-content mini-tests (either letter, number, or arrow). Mini-tests consisted of 16 items (item instantiations based on 16 distinct PDL patterns) plus two attention check items (total 18 items per mini-test). There were five mini-tests (five instantiations) for each content area (a total of 15 mini-tests, L1–L5, A1–A5, N1–N5). We created 15 forms, using a balanced incomplete blocks design with each form comprising three mini-tests in either two or three content areas (e.g., a form with L1 + L2 + A1; or a form with L1 + A1 + N1). Each participant was administered one form that consisted of 54 items altogether.

The supplemental study was designed to provide better NS items, so in it we presented only NS items. It comprised three 20-item NS mini-tests (based on 18 PDL patterns and two attention check items) distributed across four forms. One mini-test included a subset of the NS items from the main study (Study 1) that served as items to link the data from the main and supplemental studies.

Procedure

After registering, respondents linked to the online testing site and were randomly assigned one of 15 (54-item) forms. The mini-tests consisted of either LS, NS, or AS items. For each of the three mini-tests within a form, respondents were given instructions, two or three sample items, and the rules governing the series. In each test there were six simple items that served as attention checks (e.g., for LS, A B C D E F __). Participants who failed two of the attention check items were excluded from the sample by the participant panel supplier and were not counted toward the number of participants.

On average, the session took about 40 min; participants who completed the session in under 20 min were eliminated from the sample (this limit was originally set to under 30 min, but it was lowered after the first 872 participants because too many were screened out).

Results

Missing data treatments

Missing item responses were not permitted by the software, except for timing out due to item time limits (60 s for all items). Fewer than 1% of responses were in this category, and they were coded as missing in the data file (and then treated in a pairwise or casewise manner, depending on the analysis). For the purposes of computing marginal reliability, a planned missing data analysis was conducted based on the incomplete block design. Specifically, a multiple-imputation (MI) approach was used, yielding ten datasets. These datasets were analyzed separately, and parameter values were obtained as the average across the ten analyses (this is described below in the Reliability section below).

Removing items on the basis of descriptive statistics

We first computed proportions correct (P+) and point-biserial correlations (r_{bs}) for all items. The mean P+ values were .56, .44, and .09 for LS, AS, and NS, respectively (see Table 1). The difficulty of the NS items led us to also develop the supplemental study, which added new items, raising the P+ mean to .14. The mean r_{bs} values were .47, .32, .25, and .42 for LS, AS, NS (main), and NS (supplemental), respectively. We then excluded all items with P+ values less than .05 or r_{bs} less than .10 and conducted further analyses only on the remaining items. We also excluded items on the basis of response time, using the procedures described next.

Removing items on the basis of timing

Study 1 (main study) set item time limits at 1 min for all items, by giving a warning to complete at that time. There was no separate test time limit. Table 2 shows the proportions of individuals who failed to complete each item on the tests under the 1-min time limit. It can be seen that on average, 98%–99% of test takers completed all the LS and AS items, but half the test takers failed to complete the NS items. This resulted in a rewriting of the NS items for the supplemental study. In the supplemental study, the per-item time limit for NS items was raised to 2 min. These revisions resulted in completion rates similar to those found with LS and AS. For all subsequent analyses, items were excluded if fewer than 75% of the examinees were able to complete them in less than 60 s for LS and AS, 90 s for NS, and 120 s for figural matrix items (Study 2).

Table 1 Descriptive statistics (Study 1 and Study 2)

Test	Statistic	Min	Q1	Median	Q3	Max	Mean	*SD*
Letter Series (Study 1)								
	P+	.24	.45	.57	.68	.86	.56	.16
	r_{bs}	.20	.41	.47	.52	.61	.47	.09
Arrow Series (Study 1)								
	P+	.09	.30	.44	.60	.88	.44	.20
	r_{bs}	.01	.26	.33	.39	.51	.32	.10
Number Series (main)								
	P+	.00	.03	.05	.13	.54	.09	.09
	r_{bs}	– .03	.10	.21	.39	.66	.25	.19
Number Series (supplemental)								
	P+	.00	.07	.11	.17	.71	.14	.12
	r_{bs}	.12	.24	.40	.59	.76	.42	.19
Figural Matrix (Study 2)								
	P+	.06	.23	.34	.47	.78	.36	.17
	r_{bs}	.07	.34	.40	.48	.60	.40	.10

P+ = proportions correct; r_{bs} = point-biserial correlations; Q1 = 25th percentile; Q3 = 75th percentile.

Table 2 Fail-to-complete statistics by test (Study 1)

	Min	Q1	Median	Mean	Q3	Max
LS	.000	.005	.010	.012	.013	.030
AS	.000	.013	.013	.016	.018	.035
NS (main)	.000	.503	.670	.504	.672	.674
NS (supplemental)	.000	.000	.023	.023	.045	.045

Values indicate the proportion of test takers failing to complete items on the test. Thus, for LS, the mean across items was .012 failing to complete, but for the worst item, .030 failed to complete. AS = arrow series; NS = number series; LS = letter series; FM = figural matrices; Q1 = 25th percentile; Q3 = 75th percentile.

Establishing item and test time limits for future use

One purpose of this study was to establish item- and test-level time limits for the final, fixed test forms, to be used in Study 3. Because of the planned missing design used in Study 1, no individual actually took the final assembled forms, and thus it was not possible to identify the total time required to complete them. An expected total time was computed on the basis of the item response times. First, the response times for various quantiles (e.g., 75th percentile) were determined for each item. Figure 1 presents boxplots showing these values for each item in each test. Summing the response times across items at a given percentile (e.g., the 75th percentile) provided an expected response time for an examinee who consistently responded to each item in the same relative amount of time (e.g., at the 75th percentile of the item response times). If item response times across examinees were perfectly correlated, this would be an appropriate estimate of an examinee's total time. If response times across items were completely independent, then a better estimate would be the average (e.g., closer to the 50th percentile). The correlation in fact was $r = .29$, and thus the 75th percentile was likely a conservative (i.e., high) estimate of the total time required to finish the test (i.e., the actual 75th percentile on an intact form was likely to be a lower elapsed time value).

Dimensionality

We conducted principal components analyses (PCA) of item responses for each test (LS, AS, NS) separately and jointly, from the tetrachoric correlation matrices. For all three tests, scree plots suggested a large first component, and additionally one or two minor dimensions. We computed Akaike and Bayesian information criteria (AIC and BIC, respectively) fit statistics for each of the models, shown in Table 3. Whereas

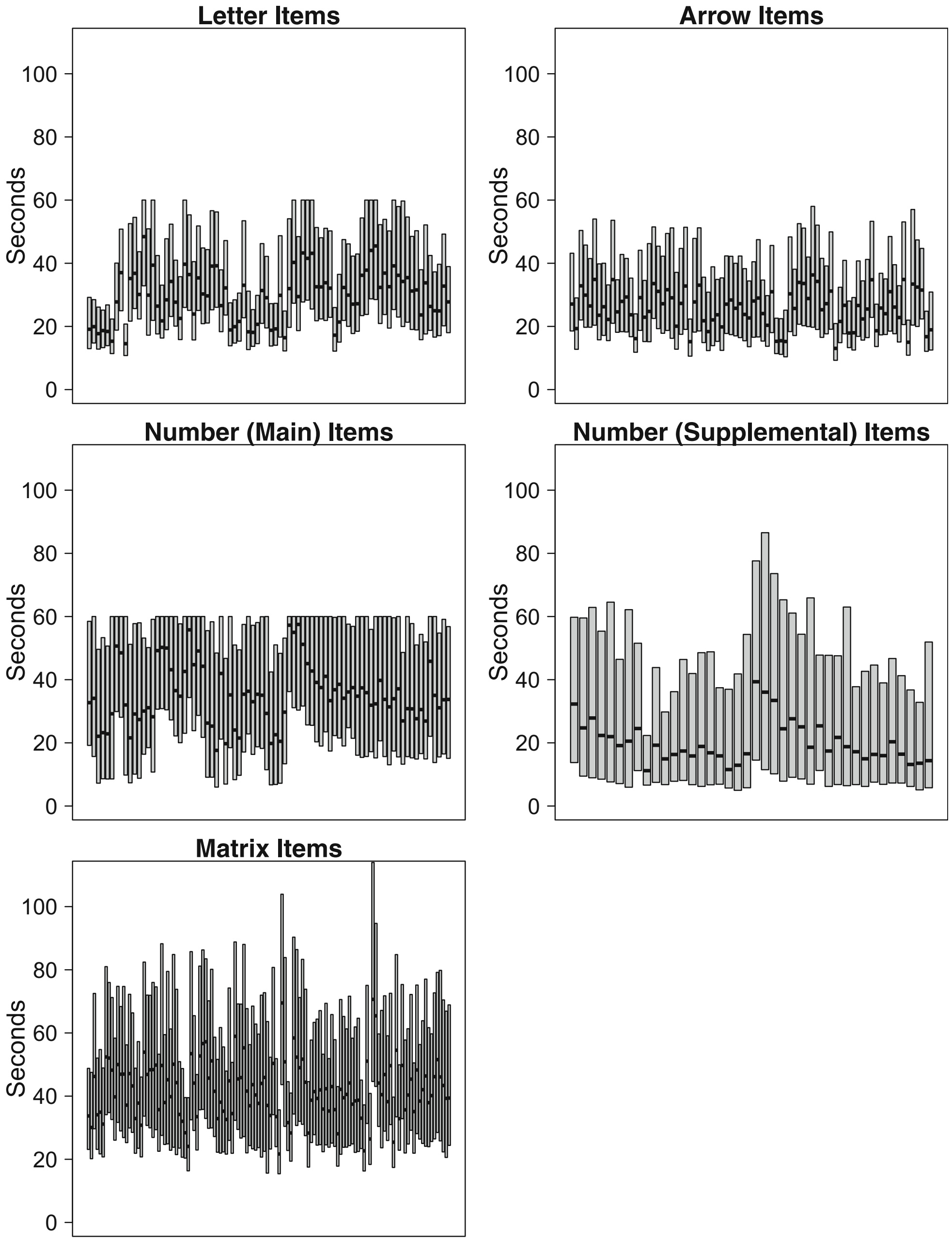

Fig. 1 Boxplots (mean with 25th and 75th percentiles) of item response times (in seconds) for Studies 1 and 2

AIC suggested two-factor solutions, BIC suggested one factor for LS, AS, and perhaps NS. For the combined set, there also was one major dimension and two minor dimensions. Several confirmatory factor analyses (CFA) were also fit to the item data: (a) a correlated primary factors model (LS, AS, NS), and (b) a bifactor model with an additional general factor (g, LS, AS, NS), where g was constrained to be independent of the correlated primary factors (for both models, means were set to zero and factor variances set to one; this assumes examinees taking each form are randomly equivalent). For the primary abilities analysis, test intercorrelations were .77, .61, and .57, for (LS, AS), (LS, NS), and (AS, NS), respectively. The test intercorrelations for the bifactor analysis were lower, at .47, .31, and .19, respectively. BIC fit statistics suggested that the

Table 3 Fit statistics (Study 1 and Study 2)

Statistic	Model	LS	AS	NS	FM
AIC	1 factor	32,766	35,090	37,893	80,925
	2 factor	32,702	34,915	37,199	80,866
BIC	1 factor	33,610	35,933	39,279	82,138
	2 factor	33,958	36,169	39,277	82,673

LS = letter series; AS = arrow series; NS = number series; FM = figural matrix (Study 2) bifactor model provided a better fit than the primary abilities model (74,723 vs. 74,027).

IRT analysis

We fit a two-parameter logistic (2PL) item response theory (IRT) model (Birnbaum, 1968) to the item responses for the remaining items. We did this separately for LS and AS, and using marginal maximum likelihood estimation via a multigroup extension of the 2PL (Bock & Zimowski, 1997) using the software program IRTPRO (Cai, Thissen, & du Toit, 2011). For the NS category, items from both the Study 1 main study and the supplemental study were calibrated concurrently using the same multigroup approach. These analyses resulted in an item bank with a final set of items and their parameters (discrimination, *a*, and difficulty, *b*).

Test forms using mixed-integer automated test assembly (ATA)

To assemble Test Forms I and II, we used a mixed-integer automated test assembly (ATA) approach (Diao & van der Linden, 2011; van der Linden, 2005). The mixed-integer linear-programming (MILP) solver lp_solve version 5.5 was called from the statistical language R (R Core Team, 2016) using the lpSolveAPI interface (Konis, 2016). Forms were assembled for each domain separately (LS, AS, NS), and then a subset of items identified for each domain was used to assemble a composite test (C). Forms were assembled using the item bank from the IRT estimation. Extremely high-difficulty items (> 5 logits) were excluded from the pool of candidate items. ATA then proceeded by identifying two sets of items (to populate Forms I and II) with equivalent test information functions (TIF) (to ensure comparable reliabilities) and test characteristic curves (TCC) (to ensure comparable difficulty and discrimination). To achieve a reliability of > .90, for the letter and arrow skill areas 30 unique items were identified for each form; for number series 30 unique and five common items were identified (35 items total) for each form.

To create parallel composite (C) tests, item parameters were first reestimated using the combined set of LS, AS, and NS items identified via ATA in the previous stage, to place all items regardless of domain on a common scale. There was some multidimensionality in the items when they were modeled together; however, items with low discrimination or high difficulty on the composite scale were excluded from the pool of candidate items. The same ATA approach described above was then used to create the composite tests, with the added constraint that 15 letter, 15 arrow, and 10 number items be included on each C form.

Reliability

We used multiple imputation via the EM algorithm to generate ten sets of item responses for the combined (final) set of letter, arrow, and number items. The imputation was done using the Amelia II package in R (Honaker, King, & Blackwell, 2011). The imputed data were used to estimate the expected abilities for each examinee (on the tests for which they took items) and to compute expected marginal reliabilities.

Scaling

With the 2PL, there is not a one-to-one correspondence between observed scores and scale scores. We obtained scale scores using Thissen and Orlando's (2001) expected a posteriori (EAP) scoring method using IRTPRO. To make the interpretation of the scale scores more intuitive, we transformed them to a *z* score metric (with associated percentiles). The mean for each scale was set to correspond with expected performance for first-year college students (based on expected change in SAT scores from first to fourth year, see Liu, 2011). Since the mean does not correspond to the empirical mean, this is a modified *z* score. Figure 2 presents the score distributions.

Validity

We computed polyserial correlations between LS, AS, NS, and education (.07, .07, .06), and Pearson correlations between Gf scores and self-reported SAT verbal (.11, .11, .10), SAT math (.24, .24, .23), and ACT (.26, .27, .28) scores. Although these correlations were lower than might have been expected, it is possible that this may at least partly have been due to the range restriction of the highly selected sample and to the unreliability of self-reports of SAT and ACT scores.

Discussion

We found that using series-generation rules from the information-processing literature, and employing item response theory and a planned missing data design, it was possible to develop parallel test forms that achieved psychometric targets of high reliability (greater than .90), good measurement across a wide range of ability from the expected first-year college student to 3.5 standard deviations above that level, essential unidimensionality, and validity evidence supporting the use of the developed measures as indicators of Gf. We also developed a procedure for setting item time limits that would make an unspeeded test, using response time percentiles. The calibrated item banks and test forms derived from our analyses here were validated in Study 3, with a sample that was administered intact forms along with an additional marker Gf test.

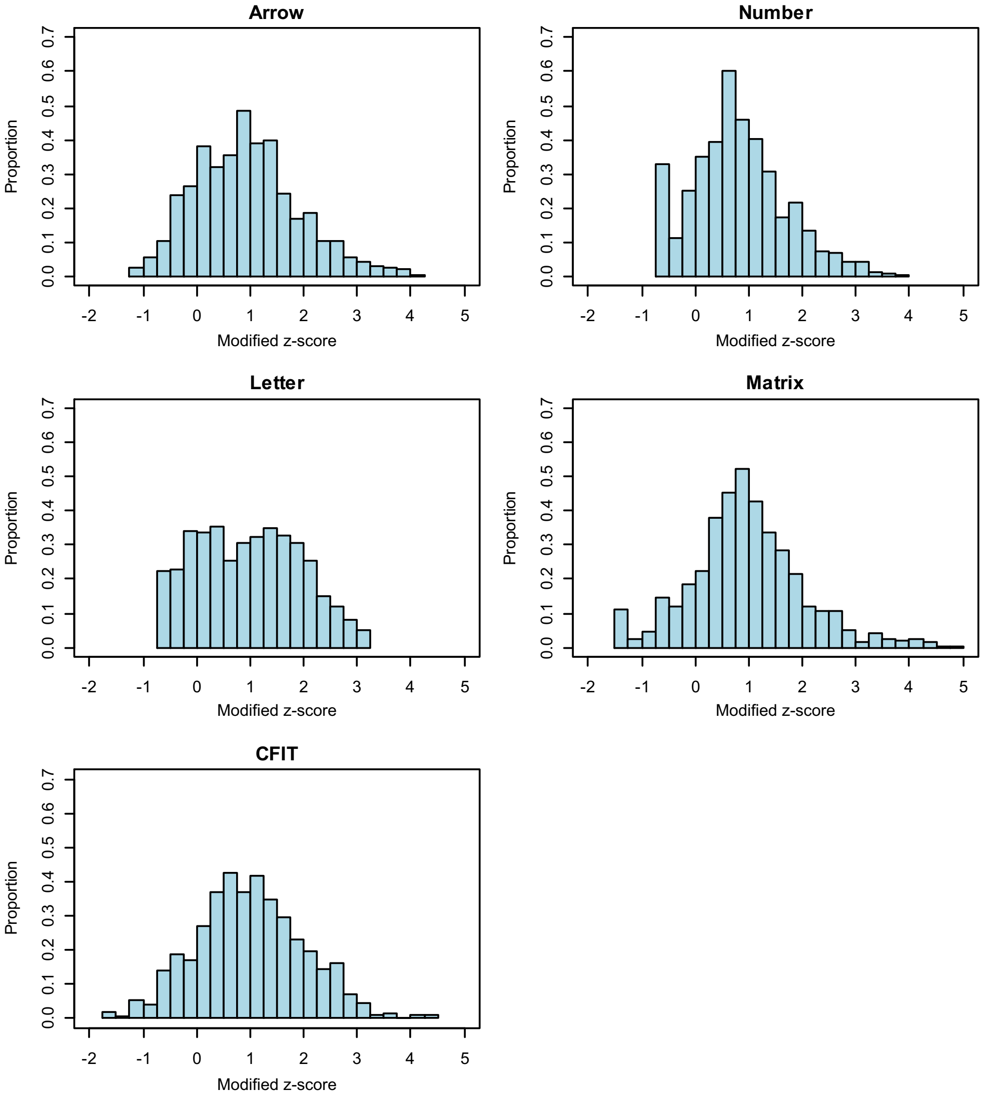

Fig. 2 Score distributions for Study 1–3 tests

Study 2: Figural matrix test development

Progressive matrices, particularly Raven's Progressive Matrices, has long been considered one of the best measures of general cognitive ability (Carpenter, Just, & Shell, 1990; Carroll, 1993; Diehl, 2002; Embretson, 2002; Snow, Kyllonen, & Marshalek, 1984). Matrix tests appear in the WAIS-III and WAIS-IV IQ tests, the most widely used individually administered intelligence tests today.

Information-processing analyses of the Progressive Matrices tests (Carpenter et al., 1990; Diehl, 2002; Embretson, 1998, 2002) have suggested the following components, which are similar to the series item models: (a) find correspondences across columns or rows (e.g., find elements, such as shapes or bars, across columns that are constant or systematically varying); (b) compare adjacent corresponding elements; and then (c) induce the element transformation rules (e.g., identity, rotation, size change) based on similarities and differences between these adjacent corresponding elements. Problems thus can be made more difficult by making element identification difficult and by increasing the number of elements or rules necessary to keep track of in working memory. For this study, we manipulated these factors in order to create a set of figural matrices items.

Our figural matrix items presented eight options in a multiple-choice format. One option was correct, and the other seven were created as variants on the correct option, with either one element or one transformation differing from the correct option.

As in Study 1, and employing a similar strategy, our goal was to create two parallel forms of items. Each test form was to be of a length sufficient for a reliability of ≥ .90, which we estimated would be about 30 items.

Method

The study was conducted in two waves (Studies 2a and 2b). The first wave was designed to check on the appropriateness of the item difficulties; the second wave excluded items with poor psychometrics characteristics, and introduced a set of new items.

Participants and measures

All participants from Study 2 were recruited from Qualtrics Labs, Inc., using a procedure similar to Study 1's. We recruited samples of 499 and 2,012 test takers for Studies 2a and 2b, respectively.

For Study 2a, we initially wrote 122 matrix items varying in expected difficulty, from 20 item models (i.e., six items per item model) and a balance of graphical families. All tasks in Study 2 were programmed for the Qualtrics Research Suite survey platform.

The 20 item models were based on the following four sets of rules: (a) addition of elements across rows and columns, (b) rotation of elements clockwise or counterclockwise, (c) position changes of elements within a given cell of a matrix, and (d) distribution of elements across rows and columns of a matrix. Rules were applied both horizontally (i.e., from the left column to the right column) and vertically (i.e., from top row to the bottom row). Up to three rules could be combined in any given item, and these rule combinations defined an item model. To create different items within an item model family, various graphical elements such as circles, triangles, and segmented lines were used; these graphical elements in various combinations were referred to as graphical families. Items were developed by applying one of 26 unique graphical families to one of the 20 item models. Response alternatives were created by altering a rule applied to one or more of the elements in the solution figure (e.g., addition vs. rotation; clockwise vs. counterclockwise rotation).

The item developers and reviewers were different authors of this article. From a careful review by reviewers (not the authors), a subset of 100 items was selected for administration for Study 2a. Items were divided into four forms designed to be equivalent with respect to theoretical difficulty and the representation of item rules and graphical families

Following Study 2a data collection (N = 499), 20 items were excluded because they were too easy or too hard (P+ < .05 or P+ > .95), had poor discrimination (r_{bs} < .10), or had a median response time of more than 1 min. New items were written to replace these and were administered with the 80 retained Study 2a items in Study 2b (a total of 100 items). The item design for Studies 2a and 2b was the same, which allowed for pooling of the administrations in order to estimate item parameters. The 80 items retained from Study 2a were administered in the same position on each of the four forms. The newly developed Study 2b items were located in the positions originally occupied by the problematic items; the new items were written and positioned in order to maintain equivalence in the theoretical difficulty and item rule and graphical family representation.

Design

Items were assembled into four forms, each with 33 items: two attention check items, 21 unique items, and ten items that were administered on one other form. There were 20 individual common items across all forms, which appeared in the same position.

Procedure

The registration and administration procedures were similar to those in Study 1. For both Studies 2a and 2b, each examinee completed one randomly assigned form. All the participants were given 55 min to complete the test, followed by a demographic questionnaire and a postsurvey asking about their on-task motivation.

Results and discussion

We used the same missing data treatment as used in Study 1. We also conducted a PCA of the item responses, based on tetrachoric correlation matrices. A scree plot indicated two or three dimensions, but examination of the item loadings did not suggest any consistent relationship with item rules or graphical families. As in Study 1, we fit the item data and estimated item parameters using marginal maximum likelihood estimation via a multigroup extension of the 2PL (Bock & Zimowski, 1997) using the software program IRTPRO (Cai et al., 2011). We used procedures similar to those in Study 1 to create two assembled forms of progressive matrices items.

We conducted a reliability analysis using the same approach as in Study 1, except that reliability was only computed for each form. The expected marginal reliability for each of the assembled forms was .98. We followed the same procedure for developing score scales as was used in Study 1.

Validity

The correlations (with standard errors of the correlations) between the figural matrices and educational attainment, self-reported SAT Verbal, SAT Math, and ACT scores were – .01 (.02), .03 (.03), .13 (.03), and .07 (.03), respectively, which are slightly lower than the ones we found in Study 1.

Establishing item and test time limits for future use

For Study 2, we increased the item time limit to 2 min, due to the complexity of figural matrix items as compared to the series items. We used the same procedure as in Study 1 to set item-specific time limits in preparation for Study 3. The expected total response times were approximately 21, 32, and 48 min (all responses), or 23, 32, and 43 min (correct responses only), at the 50th, 75th, and 90th completion-time percentiles, respectively. That is, 90% of the examinees would be expected to complete the test in 48 min (either form), and 75% in 32 min or less. A standard criterion for establishing that a test is unspeeded is that virtually all examinees complete at least three-fourths of the test (e.g., Rindler, 1979; Swineford, 1974). Assuming time needed to complete the test is linear with the percentile then using the methodology we used to establish a total test time limit based on the summed 75th per item time percentiles achieves the unspeeded criterion.

Assembling parallel forms with certain psychometric properties and suggested time limits as we did here was based on statistical estimates, because no one participant actually took all the figural matrix items developed. In Study 3 we assembled parallel fixed forms based on the results from Studies 1 and 2 and administered these forms to other participants.

Study 3: Validation study

Studies 1 and 2 involved a planned missing design, and only statistical estimates of item parameters and timing information could be made. In Study 3 we prepared fixed test forms to verify that the estimated statistical and psychometric parameters would hold when items were administered to a common sample. A second purpose of this study was to administer an independent marker test of Gf, Cattell's (1949; Cattell, Krug, & Barton, 1973) Culture Fair Intelligence Test, Form 3 (CFIT), and to compute correlations with scores on that measure.

Method

We followed Studies 1 and 2 in targeting an online panel of 800 test takers for data collection. Approximately 100 were eliminated due to not completing the tests or to completing the tests too quickly, but recruiting continued until the target of 800 was met.

On the basis of the results from Studies 1 and 2, we assembled test forms. We also administered a standard Gf marker test, Form 3 of the CFIT (Cattell et al., 1973). Forms differed by varying the order in which the figural matrices (FM), number series (NS), letter series (LS), arrow series (AS), and CFIT tests were administered.

Test time limits were set on the basis of the time-limit estimation procedures implemented in Studies 1 and 2. The overall time limit was approximately 180 min, with 30 min each for the letter (LS) and arrow series (AS), 45 min each for number series (NS) and figural matrices (FM), and 12.5 min (the manual specified time limit) for the CFIT. Item-specific time limits were also imposed as follows: 1 min for LS and AS items, 1.5 min for NS items, and 2 min for FM items. As part of the administration, test takers were given the opportunity to take breaks between tests; however, we imposed forced breaks of 5 min after every second test. As a final component of the study design, we used attention check items in combination with a timing threshold of 23 min per form (equal to 1/3 of the median expected total time) to exclude participants who provided too rapid responses (these examinees were eliminated from the sample by the participant panel supplier prior to any analysis, and they were not included in the sample count).

Prior to taking the items for a given test, examinees were given instructions and a few example items of low to average difficulty. At the end of the assessment, they were asked to complete a background questionnaire that included demographic information and examinees' SAT, ACT, and GRE scores as in Studies 1 and 2.

Results

Item analysis

No items from the FM, NS, LS, or AS tests were identified as being too hard or too easy (i.e., having extremely low or high P+ values below .05 or above .95). Nor did any of the items have extremely low or negative biserial correlations. However, four of the CFIT items were too hard (i.e., P+ < .05), and an additional three of these items had negative biserial correlations. All seven of these items were removed prior to further analyses.

To evaluate whether the statistical properties of the tests obtained in the present study were consistent with those found in Studies 1 and 2, we conducted an evaluation of item fit. We estimated item parameters for each of the four tests using the data collected for the present studies, then used the Stocking–Lord (1983) method to place the items and the ability estimates onto the previously established scales. The linking coefficients obtained in this process showed that the mean abilities for Study 3 on each test were lower (and less variable) than the means for Studies 1 and 2. However, none of the transformed item parameter estimates differed significantly from those estimated in the previous studies. We also found no evidence for fatigue effects by conducting analyses of variance on percentages correct for the five test positions.

Reliability

The observed IRT marginal reliabilities for the tests were .89, .89, .95, and .88 for AS, NS, LS, and FM, respectively (the CFIT reliability was .92). For NS and LS, the observed reliabilities were only slightly lower than the expected reliabilities, based on Study 1 (.90 and .96, respectively); however, for AS and FM, the differences between the observed and expected reliabilities were notably larger (.95 and .98, respectively), for unknown reasons.

Dimensionality

A PCA suggested that there was one large and several minor components. A unidimensional model, two EFA models, and three CFA models were also fit for the combined set of items. One CFA model was a simple structure model in which the items from each test loaded on a separate dimension. One bifactor model included a general factor and specific factors for each of the five tests. Another bifactor model included two specific factors. For all of the models, the means were constrained to be zero, and the factor variances for all groups were constrained to unity.

On the basis of BIC fit statistics, the worst-fitting model was the simple structure model, followed by the five-specific bifactor model (see Table 4). The unidimensional model fit better than the EFA two-factor model but did not appreciably differ from the EFA three-factor model. An examination of the loadings from the EFA two-factor model showed that the arrow, matrix, and CFIT items loaded predominantly on a single factor, perhaps reflecting their common spatial content, whereas the number and letter items loaded predominantly on the other factor. On the basis of these findings, the two specific factors in the two-specific bifactor model were (a) NS and LS items loading on one and (b) AS, FM, and CFIT items on the other. The two-specific bifactor model was the best-fitting model. These results suggest a single dominant general factor that extended across all five tests, but that spatial versus nonspatial dimensions were associated with the tests, as well.

Table 4 Model fits for all tests combined (Study 3)

Model	BIC
Unidimensional	133,036
EFA – 2 factors	133,802
EFA – 3 factors	132,964
Simple structure	145,424
Bifactor – 5-specific	134,740
Bifactor – 2-specific	132,500

Timing information

Table 5 suggests that the majority of examinees were able to respond to the items in a reasonable amount of time. Note that response times were not collected for individual items on the CFIT; items were administered in item sets in order to maintain consistency with how the test is typically administered in the paper-and-pencil format.

We also noted some extremely short minimum times. We encouraged examinees to provide adequate effort; for example, a very fast test completion time (AS < 4.5 min, NS < 5 min, LS < 5 min, FM < 7 min) would be followed by a warning, and successive warnings led to removal from the testing session. The minimum times in Table 5 correspond to these fast responders. No data were available on the CFIT regarding response times prior to this study, so no minimum times were specified for CFIT.

Test level descriptive statistics

Using the item parameters from Studies 1 and 2, EAP scores were estimated for each examinee on each of the tests (new item parameters for the CFIT were estimated using the 2PL). These values are reported on a modified *z*-score metric in which the expected performance for first-year college students was set to 0 (*SD* = 1), for interpretability (Table 6).

Table 7 presents disaggregated results by demographics. Because the data were scaled so that the mean standard deviation was 1, it was possible to compute approximate effect sizes as the difference between the subgroup means.

Relationship to external validation measures

All tests correlated strongly with the CFIT (Table 8), providing strong evidence of convergent validity and supporting the tests developed here as measures of Gf. We also computed correlations of the Gf measures with educational attainment (polyserial correlations) and with SAT and ACT scores (Pearson correlations). It can be seen that the correlations

Table 5 Item completion times (in seconds) (Study 3)

Percentile	AS		NS		LS		FM	
	Low	High	Low	High	Low	High	Low	High
5th	2.34	7.58	3.49	16.65	2.73	9.48	2.25	7.17
50th	10.59	24.82	15.57	60.95	11.59	43.35	14.34	46.68
75th	15.61	39.33	30.89	89.04	16.34	59.99	24.81	73.2
90th	23.99	56.42	56.47	90.01	24.35	60.01	41.36	104.16

Values are the range (low, high) of item completion times for different items at various completion time percentiles. Thus, low values are low time intensity items. AS = arrow series; NS = number series; LS = letter series; FM = figural matrices.

Table 6 Scale scores for Gf tests (Study 3)

	Min	Q1	Median	Mean	Q3	Max
AS	– 1.22	0.20	0.87	0.93	1.49	4.03
NS	– 0.74	0.22	0.73	0.81	1.32	3.76
LS	– 0.71	0.16	0.97	0.96	1.72	3.23
FM	– 1.43	0.37	0.90	0.95	1.52	4.89
CFIT	– 1.56	0.34	0.98	1.00	1.66	4.37

Test scores scaled to a mean of 0 for expected performance of a 1st-year college student (*SD* = 1). Min = lowest score by an individual; Q1 = 25th percentile; Q3 = 75th percentile; Max = highest score by an individual. AS = arrow series; NS = number series; LS = letter series; FM = figural matrices; CFIT = Culture Fair Intelligence Test.

between Gf and other measures (a) were not that high in general, but (b) were generally highest for SAT Quantitative and for education level.

General discussion

Research on enhancing fluid intelligence has grown in popularity recently. However, many current efforts suffer from weaknesses in how fluid ability is measured. The purpose of this study was to develop a battery of Gf tests that could be used in studies designed to improve fluid ability. For this purpose, we designed a battery that was on average reliable, would provide precise measurements across a wide range of abilities, offer true parallel forms for pre–post administration, include multiple types of tests that sampled from a range of content domains and methods, and provide time limits that would enable the measurement of level of ability not confounded by speededness. We employed item response theory methods to create a test battery specifically designed for pre- and post-intervention administration in intervention-focused studies. The results, based on administering items to over 6,300 highly educated individuals, ranging from current upperclassmen to Ph.D. holders, showed that the battery consisting of four different tests was highly reliable (each test had a reliability of > .90) and yielded clear evidence for the battery's construct validity as a measure of Gf. As such, the Gf battery represents a promising tool for measuring the effectiveness of cognitive enhancement interventions.

Table 7 Test scale scores by subgroup (Study 3)

		Letter	Arrow	Number	FM	CFIT
Gender	Male	0.96	0.98	0.88	0.97	0.97
	Female	0.95	0.88	0.73	0.92	1.03
Education	3rd/4th Yr. College	0.77	0.87	0.58	0.89	0.97
	College Graduate	0.79	0.78	0.62	0.75	0.76
	Masters Student	0.84	0.90	0.65	0.91	0.90
	Master's Degree	1.16	0.99	1.00	1.10	1.11
	Doctorate Degree	1.22	1.11	1.19	1.10	1.24
Age	18–25	0.86	1.01	0.65	1.01	1.05
	26–40	0.85	0.82	0.72	0.83	0.80
	41–60	1.14	1.05	0.99	1.06	1.08
	61 and over	1.13	0.88	1.03	1.01	1.18

Within-cell standard deviations ranged from 0.70 to 1.19, with a mean of 1.0. FM = figural matrices; CFIT = Culture Fair Intelligence Test.

Table 8 Correlations among Gf tests, education level, and college admissions test scores (Study 3)

	AS	NS	LS	FM	CFIT
Education level	.09	.25	.18	.10	.12
SAT Quantitative	.14	.17	.16	.12	.03
SAT Verbal	– .07	– .03	– .03	.01	– .09
ACT	– .02	.07	.00	.01	– .05
AS		.64	.74	.72	.73
NS	.57		.66	.64	.54
LS	.68	.61		.64	.66
FM	.64	.57	.59		.64
CFIT	.66	.49	.61	.58	

Disattenuated correlations among Gf scores are reported in the upper triangle. AS = arrow series; NS = number series; LS = letter series; FM = figural matrices; CFIT = Culture Fair Intelligence Test.

We developed four tests based on information-processing models from the literature. In Studies 1 and 2, based on data from 4,500 college students and advanced graduates in an online panel, we assembled two parallel forms for each test. In Study 3 we administered the four tests and an additional Gf marker test (the CFIT) in a spiraled design (in order to avoid test position effects) to 802 online panel participants equally divided across advanced education levels. We found that the four tests in the battery performed psychometrically as we had predicted from Studies 1 and 2, with reliabilities around .90, covering a range of difficulty, and having high ceilings and good measurement precision across an ability range from college students to advanced graduates and beyond.

Factor analyses indicated a strong general factor across the tests, and additionally, a spatial versus nonspatial secondary factor. This finding suggests that the battery might be useful both as a means to assess overall effects of an intervention designed to enhance general fluid/inductive reasoning ability, and also to evaluate the differential effects of an intervention designed to target the enhancement of spatial versus nonspatial abilities.

Because automatic item generation procedures were used to develop the items for this study, it would be possible to modify a targeted difficulty level in order to create a test that could be easier or more difficult than the tests in the battery assembled here. In this study we identified a number of key construct-relevant difficulty factors for each of the four tests and used those factors to generate items with predicted difficulty levels. This strategy generally was successful, as is shown by the fact that the items developed from the models achieved the targeted difficulty levels. However, additional item analyses could be conducted to identify and quantify the effects of the various difficulty factors implemented here more precisely, for the purposes of creating a test battery at a different targeted difficulty level.

In summary, renewed interest in cognitive ability enhancement—which is evident in increased scholarly activity, the presence and profitability of commercial offerings, and government investment—demands increased attention to issues of research design and measurement integrity (Shipstead et al., 2012). Previous studies have produced mixed results. In response to the psychometric shortcomings of these studies, our intention was to create a test battery specifically designed for pre- and postintervention administration in intervention-focused studies. The resulting battery could be implemented in future Gf intervention studies.

Author note This publication is based on work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) SHARP program, via contract 2015-14120200002-002, and is subject to the regulation Rights in Data—General Clause, CFR 52.227-14, Alt. IV (DEC 2007). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, expressed or implied, of ODNI, IARPA, or the US government, or of any of the authors' host affiliations. The US government is authorized to reproduce and distribute reprints for governmental purposes, notwithstanding any copyright annotations therein. Technical information on the material covered in this article, including on administration, instructions, test items, answer keys, scoring, test properties, and related content, is available in the publication *Guide to MITRE/Educational Testing Service (ETS) Inductive Reasoning: Battery for a High Ability Population*, https://www.mitre.org/publications/technical-papers/guide-to-mitre-educational-testing-service-inductive-reasoning-battery. © 2016 The MITRE Corporation, all rights reserved. Approved for public release, distribution unlimited. Case number 16-1595.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: A meta-analysis. *Psychonomic Bulletin & Review*, *22*, 366–377. https://doi.org/10.3758/s13423-014-0699-x

Basak, C., Boot, W. R., Voss, M. W., & Kramer, A. F. (2008). Can training in a real-time strategy videogame attenuate cognitive decline in older adults? *Psychology and Aging*, *23*, 765–777.

Bertling, J. P. (2012). *Measuring reasoning ability: Applications of rule-based item generation*. Unpublished doctoral dissertation, University of Münster.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinees' ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 433–448). New York, NY: Springer. https://doi.org/10.1007/978-1-4757-2691-6_25

Bridgeman, B., & Rock, D. A. (1993). Relationships among multiple-choice and open-ended analytical questions. *Journal of Educational Measurement*, *30*, 313–329.

Bridgeman, B., Trapani, C., Curley, E. (2003). Effect of fewer questions per section on SAT® I scores (College Board Research Report No. 2003-2, pp. 1–16). New York, NY: College Board. https://doi.org/10.1002/j.2333-8504.2003.tb01900.x

Buschkuehl, M., & Jaeggi, S. M. (2010). Improving intelligence: A literature review. *Swiss Medical Weekly*, *140*, 266–272.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*, 404–431.https://doi.org/10.1037/0033-295X.97.3.404

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

Cattell, R. B., Krug, S.E., & Barton, K. (1973). Technical supplement for the Culture Fair Intelligence Tests, Scales 2 and 3. Champaign, IL: Institute for Personality and Ability Testing.

Chein, J. M., & Morrison, A. B. (2010). Expanding the mind's workspace: Training and transfer effects with a complex working memory span task. *Psychonomic Bulletin & Review*, *17*, 193–199.https://doi.org/10.3758/PBR.17.2.193

Cliffordson, C., & Gustafsson, J. E. (2008). Effects of age and schooling on intellectual performance: Estimates obtained from analysis of continuous variation in age and length of schooling. *Intelligence*, *36*, 143–152.

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_Solve version 5.5 in R. *Applied Psychological Measurement*, *35*, 398–409.https://doi.org/10.1177/0146621610392211

Diehl, K. A. (2002). *Algorithmic item generation and problem solving strategies in matrix completion items*. Unpublished doctoral dissertation, University of Kansas.

Dinges, D. F., & Weaver, T. E. (2003). Effects of modafinil on sustained attention performance and quality of life in OSA patients with residual sleepiness while being treated with nCPAP. *Sleep Medicine*, *4*, 393–402.

Dougherty, M. R., Hamovitz, T., & Tidwell, J. W. (2016). Reevaluating the effectiveness of *n*-back training on transfer through the Bayesian lens: Support for the null. *Psychonomic Bulletin & Review*, *23*, 306–316.https://doi.org/10.3758/s13423-015-0865-9

Duthie, S. J., Whalley, L. J., Collins, A. R., Leaper, S., Berger, K., & Deary, I. J. (2002). Homocysteine, B vitamin status, and cognitive function in the elderly. *American Journal of Clinical Nutrition*, *75*, 908–913.

Eisen, S. V., Ranganathan, G., Seal, P., & Sprio, A. (2007). Measuring clinically meaningful change following mental health treatment. *Journal of Behavioral Health Services and Research*, *34*, 272–289.

Ekstrom, J. W., French, J. W., Harman, H. H., & Dermen, D. (1976). Manual for Kit of Factor Referenced Cognitive Tests (pp. 109–113). Princeton, NJ: Educational Testing Service.

Embretson, S. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika 56*, 495–515.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*, 380–396. https://doi.org/10.1037/1082-989X.3.3.380

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. Irvine & P. Kyllonen (Eds.), Item generation for test development (pp. 219–250). Mahwah, NJ: Erlbaum.

Emmerich, W., Enright, M. K., Rock, D. A., & Tucker, C. (1991). The development, investigation, and evaluation of new item types for the GRE analytical measure (GRE Board Professional Report No. 87-09P, ETS RR-91-16). Princeton, NJ: Educational Testing Service.

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working memory training may increase working memory capacity but no fluid intelligence. *Psychological Science*, *24*, 2409–2419. https://doi.org/10.1177/0956797613492984

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2015). Do we really become smarter when our fluid-intelligence test scores improve? *Intelligence*, *48*, 1–14. https://doi.org/10.1016/j.intell.2014.10.005

Hernstein, R. J., Nickerson, R. S., de Sánchez, M., & Swets, J. A. (1986). Teaching thinking skills. *American Psychologist*, *41*, 1279–1289. https://doi.org/10.1037/0003-066X.41.11.1279

Holzman, T. G., Pellegrino, J. W., & Glaser, R. (1983). Cognitive variables in series completion. *Journal of Educational Psychology*, *75*, 603–618.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data (R package version 1.5-5) [Computer software]. Retrieved from https://www.inside-r.org/packages/cran/Amelia

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19. https://doi.org/10.1037/0022-006X.59.1.12

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, *105*, 6829–6833. https://doi.org/10.1073/pnas.0801268105

Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. *Memory & Cognition*, *42*, 464–480. https://doi.org/10.3758/s13421-013-0364-z.

Jaeggi, S. M., Studer, B., Buschkuehl, M., Su, Y.-F., Jonides, J., & Perrig, W. J. (2010). On the relationship between N-back performance and matrix reasoning—Implications for training and transfer. *Intelligence*, *38*, 625–635.

Jäger, A. O., Holling, H., Preckel, F., Schulze, R., Vock, M., Süß, H.-M., & Beauducel, A. (2006). *Berliner Intelligenzstrukturtest für Jugendliche (BIS-HB): Begabungs- und Hochbegabungsdiagnostik.* Göttingen, Germany: Hogrefe.

Jaušovec, N., & Jaušovec, K. (2012). Working memory training: Improving intelligence—Changing brain activity. *Brain and Cognition*, *79*, 96–106. https://doi.org/10.1016/j.bandc.2012.02.007

Karbach, J., & Kray, J. (2009). How useful is executive control training? Age differences in near and far transfer of task-switching training. *Developmental Science*, *12*, 978–990.

Karbach, J., & Verhaeghen, P. (2014). Making working memory work: A meta-analysis of executive-control and working memory training in older adults. *Psychological Science*, *25*, 2027–2037.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating.* New York, NY: Springer.

Konis, K. (2016). lpSolveAPI, version 5.5.2.0 [Computer software]. Retrieved from https://CRAN.R-project.org/package=lpSolveAPI

Korol, D. L. (2002). Enhancing cognitive function across the life span. *Annals of the New York Academy of Science*, *959*, 167–179.

Kotovsky, K., & Simon, H. A. (1973). Empirical tests of a theory of human acquisition of concepts for sequential patterns. *Cognitive Psychology*, *4*, 399–424.

Kyllonen, P. C., Roberts, R., & Stankov, L. (2008). Extending intelligence: Enhancement and new constructs. Mahwah, NJ: Erlbaum.

Kyllonen, P. C., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, *4*(4), 14. https://doi.org/10.3390/jintelligence4040014

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, *41*, 352–362. https://doi.org/10.3102/0013189X12459679

McGrew, K. S. (2009). CHC theory and the Human Cognitive Abilities Project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*, 1–10. https://doi.org/10.1016/j.intell.2008.08.004

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer": Evidence from a meta-analytic review. *Perspectives on Psychological Science*, *11*, 512–534. doi: https://doi.org/10.1177/1745691616635612

MITRE/Educational Testing Service. (2016). Guide to MITRE/Educational Testing Service (ETS) Inductive Reasoning Battery for a High Ability Population. https://www.mitre.org/publications/technical-papers/guide-to-mitre-educational-testing-service-inductive-reasoning-battery.

Moody, D. E. (2009). Can intelligence be increased by training on a task of working memory? *Intelligence*, *37*, 327–328. https://doi.org/10.1016/j.intell.2009.04.005

Myors, B., Stankov, L., & Oliphant, G. W. (1989). Competing tasks, working memory and intelligence. *Australian Journal of Psychology*, *41*, 1–16.

OECD. (2017). Programme for International Student Assessment: PISA 2015 (Technical Report, Chapter 9, Scaling PISA data). Paris: Author. Retrieved from https://www.oecd.org/pisa/data/2015-technical-report/

R Core Team. (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Rae, C., Digney, A. L., McEwan, S. R., & Bates, T. C. (2003). Oral creatine monohydrate supplementation improves brain performance: A double-blind, placebo-controlled, cross-over trial. *Proceedings of the Royal Society B*, *270*, 2147–2150.

Ramos, B. P., & Arnsten, A. F. (2007). Adrenergic pharmacology and cognition: Focus on the prefrontal cortex. *Pharmacology and Therapeutics*, *113*, 523–536.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., . . . Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology: General*, *142*, 359–379. https://doi.org/10.1037/a0029082

Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement*, *16*, 261–270.

Robin, F., Steffen, M., & Liang, L. (2014). The multistage test implementation of the GRE Revised General Test. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), Computerized multistage testing: Theory and applications (pp. 325–341). Boca Raton, FL: Chapman and Hall/CRC.

San Martín, E., González, J., & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3PL model. *Psychometrika*, *80*, 450–467. https://doi.org/10.1007/s11336-014-9404-2

Schneider, W. J., & McGrew, K. S. (2012). The Cattell–Horn–Carroll model of intelligence. In D. P. Flanagan, & P. L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (pp. 99–144). New York, NY: Guilford Press.

Schwaighofer, M., Fischer, F., & Bühner, M. (2015). Does working memory training transfer? A meta-analysis including training conditions as moderators. *Educational Psychologist*, *50*, 138–166.

Sellers, K. K., Mellin, J. M., Lustenberger, C. M., Boyle, M. R., Lee, W. H., Peterchev, A. V., & Frohlich, F. (2015). Transcranial direct current stimulation of frontal cortex decreases performance on the WAIS-IV intelligence test. *Behavioural Brain Research*, *290*, 32–44. https://doi.org/10.1016/j.bbr.2015.04.031

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*, *138*, 628–654. doi: https://doi.org/10.1037/a0027473

Simon, H. A., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review*, *70*, 534–546.

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. L. (2016). Do "brain training" programs work? *Psychological Science in the Public Interest*, *17*, 103–186. https://doi.org/10.1177/1529100616661983

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 2, pp. 47–103). Hillsdale, NJ: Erlbaum.

Stankov, L. (1986). Kvashchev's experiment: Can we boost intelligence? *Intelligence*, *10*, 209–230.

Stankov, L., & Cregan, A. (1993). Quantitative and qualitative properties of an intelligence test: Series completion. *Learning and Individual Differences*, *5*, 137–169.

Stankov, L., & Myors, B. (1990). The relationship between working memory and intelligence: Regression and COSAN analyses. *Personality and Individual Differences*, *11*, 1059–1068.

Stephenson, C. L., & Halpern, D. F. (2013). Improved matrix reasoning is limited to training on tasks with a visuospatial component. *Intelligence*, *41*, 341–357. https://doi.org/10.1016/j.intell.2013.05.006

Sternberg, R. J. (2008). Increasing fluid intelligence is possible after all. *Proceedings of the National Academy of Sciences*, *105*, 6791–6792.

Stine-Morrow, E. A. L., Parisi, J. M., & Morrow, D. G. (2008). The effects of an engaged lifestyle on cognitive vitality: A field experiment. *Psychology and Aging*, *23*, 778–786.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589–617.

Swineford, F. (1974). The test analysis manual (ETS SR 74-06). Princeton, NJ: Educational Testing Service.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds), Test scoring (pp. 73–140). Hillsdale, NJ: Erlbaum.

Thorndike, R. L., & Hagen, E. (1971). Cognitive abilities tests. Boston, MA: Houghton Mifflin.

Thurstone, L. L., & Thurstone, T. G. (1941). Factorial studies of intelligence (Psychometric Monographs, No. 2). Chicago, IL: University of Chicago Press.

Tranter, L. J., & Koutstaal, W. (2008). Age and flexible thinking: An experimental demonstration of the beneficial effects of increased cognitively stimulating activity on fluid intelligence in healthy older adults. *Aging, Neuropsychology, and Cognition*, *15*, 184–207.

Valentin Kvist, A., & Gustafsson, J.-E. (2008). The relation between fluid intelligence and the general factor as a function of cultural background: A test of Cattell's Investment theory. *Intelligence*, *36*, 422–436. https://doi.org/10.1016/j.intell.2007.08.004

van der Linden, W. J. (2005). Linear models for optimal test design. New York, NY: Springer.

von Bastian, C. C., & Oberauer, K. (2013). Distinct transfer effects of training different facets of working memory capacity. *Journal of Memory and Language*, *69*, 36–58.

Weeks, J. P., Kyllonen, P. C., Bertling, M., & Bertling, J. P. (2017). Technical supplement to general fluid/inductive reasoning battery for a high-ability population. Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Wilhelm, O. (2005). Measuring reasoning ability. In O. Wilhelm & R. W. Engle (Eds.), Handbook of understanding and measuring intelligence (pp. 373–392). Thousand Oaks, CA: Sage.