Measuring Cognitive Aptitude Using Unobtrusive Knowledge Tests: A New Survey Technology

Peter J. Legree
Daniel E. Martin
Joseph Psotka
U.S. Army Research Institute, Alexandria, VA, USA

Five knowledge tests and one implicit-reasoning task were developed to be: (1) exceptionally short, (2) correlated with general cognitive aptitude, (3) unobtrusive, i.e., appear similar to attitudinal survey items as opposed to maximal performance measures, and (4) without formally "correct" answers. The intent was to design scales that could be administered in non-proctored environments to directly measure general cognitive aptitude while avoiding the possibility that participants could use references to provide "good" answers. The five knowledge tests used a Likert format to assess knowledge in verbal and practical domains, and were scored by computing distances between examinee and reference ratings. The implicit-reasoning task appeared to be a series completion "game" that required a dichotomous response. The scales were administered to 288 Air Force recruits and were validated against the Armed Services Vocational Aptitude Battery (ASVAB). Individual unobtrusive knowledge scales and ASVAB tests were substantially correlated with sample correlations ranging to .39 and population correlation estimates to .66 after correcting for range restriction. Two sets of factor scores, which were separately derived from the unobtrusive test battery and the ASVAB, were highly correlated in our sample, .54, yielding a population correlation of .80 after correcting for range restriction. This technology is important because few paper- or Internet-based surveys, and virtually no mail-based surveys accurately measure general cognitive aptitude, while many of these surveys address important social issues and commercial questions that could be better understood given an unobtrusive but accurate estimate of general cognitive aptitude.

In earlier research, we demonstrated that accurate cognitive aptitude scores could be obtained by administering a computer adaptive test of word knowledge over the telephone (Legree, Fischl, Gade, & Wilson, 1998). The adaptive test required less than 10 min to administer and we speculated that the procedure could provide valuable social and psychological insights if it were incorporated into a more general survey. That accomplishment was important because while many surveys address important issues, no practical method existed to accurately estimate general cognitive aptitude in those surveys.

Direct all correspondence to: Peter J. Legree, U.S. Army Research Institute, 5001 Eisenhower Avenue, Alexandria, VA 22333-5600, USA. E-mail: legree@ari.army.mil

INTELLIGENCE 28(4): 291–308
ISSN: 0160-2896
Published by Elsevier Science Inc.

Unfortunately, the telephone testing approach uses a computer adaptive test and is limited in applicability to computer-assisted telephone interviews, while many more surveys are mail-, Internet-, or paper-based. An additional problem associated with the telephone testing procedure is that administering a computer adaptive test over the telephone requires the permission of the agency that underwrote the test development effort and that agency may not want to risk its investment.

The current project was designed to extend our earlier findings by developing and validating a battery of inexpensive cognitive scales that could be incorporated into mail-, Internet-, or paper-based surveys and used to accurately measure general cognitive aptitude in a much larger number and variety of applications.[1]

Special Constraints Associated With Mail, Internet, or Paper Surveys

There are at least two important constraints that must be considered when developing a scale to estimate cognitive aptitude within a mail- or paper-based survey. First, mail- and Internet-based surveys are not proctored and test control is minimal for these questionnaires; once the instruments are distributed, participants may or may not return or duplicate them. Second, return rates are inversely proportional to the length of the questionnaire; therefore, any procedure used to estimate cognitive aptitude must be highly efficient to allow other data to be collected.

Test Control

We believe that it is not reasonable to include a conventional cognitive test in mail-based or non-proctored surveys because of the possibility that some subjects will consult reference sources to provide "correct" answers and thereby invalidate the test scores. While this statement reflects our opinion, it is relevant that we do not know of any mail survey that has included a cognitive aptitude test. While directly measuring cognitive aptitude in a mail survey may appear to be an intractable problem, we decided to try to develop knowledge tests that: (1) were not and did not appear to be conventional knowledge tests, and (2) tapped general knowledge domains or skills for which standard references were not available.

To accomplish this goal, we investigated the possibility of developing Likert knowledge tests to measure general cognitive aptitude. Likert *knowledge* tests require participants to rate a set of items on a common scale to demonstrate expertise in some knowledge domain and are therefore different from most Likert scales that are developed for attitudinal or opinion research. For example, a Likert knowledge test might present a social problem and require subjects to rate the relative appropriateness (common scale) of 20 possible actions (items). Performance is scored for each item as the distance between a respondent's rating and a reference value with smaller distances indicating better performance (cf. Legree, 1995). The reference value represents the average rating provided by a representative group of individuals for that item.

There are two important advantages associated with developing Likert knowledge tests to assess general cognitive aptitude for mail, Internet, or paper surveys. First, standard reference sources (e.g., dictionaries and encyclopedias) do not address the domains assessed by the Likert tests, and this characteristic makes it difficult for a participant to consult a reference source to provide “good” answers, i.e., cheat. Second, these scales often appear to be assessing opinions or attitudes, as opposed to knowledge, and this characteristic lowers the likelihood that many participants would attempt to find a reference source to provide “good” answers. It is relevant that the senior author had developed Likert knowledge tests for other domains and was struck by the difficulty of convincing individuals, including research psychologists, that the scales constituted maximal-performance measures.

These considerations led the senior author to believe that Likert knowledge tests could be developed in which few individuals would recognize as “conventional tests” for domains for which reference sources are not available to verify answers. These scales are unobtrusive to the extent that they do not appear to evoke test anxiety or to be recognized as “tests.” To describe these scales, we coined the terms “*Unobtrusive Knowledge Test*” or “UKT.”

Survey Length

It is important to realize that space and time are worth a premium within a survey questionnaire. One unusual characteristic of Likert knowledge tests is that items can be designed to be extremely short. In the experimental tests we developed, each item was either one or two words long. Thus, each of the Likert knowledge tests contained between 15 and 30 items and required between 15 and 37 words to present the items. We know of no conventional test with a lower word-to-item ratio and it is accurate to characterize these knowledge tests as extremely efficient.

A second advantage to the Likert format is that a distance is calculated for each item, and all items contribute to an individual score. In a conventional test, items must be chosen to cover a broad range of item difficulty. As is well known, the broad range of item difficulty results in wasted effort because many items are either too easy or too difficult for any particular individual. By using a Likert format to develop knowledge scales, we believed that we could develop an experimental battery that would include scales that could be administered quickly and efficiently.

Given the nature of this application, we reasoned that Likert knowledge tests could be developed for domains corresponding to the Armed Services Vocational Aptitude Battery (ASVAB) factor structure. By developing Likert scales corresponding to multiple factors, we expected that the scales would collectively load at a substantial level on *Psychometric g*.

RESEARCH GOALS

The above considerations led us to try to develop Likert knowledge tests that would be:

1. Unobtrusive in the sense that they would not appear to be conventional tests,
2. Unlikely to be compromised,
3. Exceptionally short and efficient,
4. Correlated substantially with Psychometric *g*.

We felt that scales with the above characteristics would be suitable for mail-, Internet-, or paper-based survey administration. This article reports the validation of these tests.

Method

Our experimental package contained nine scales corresponding to: a demographic self-report sheet, the five unobtrusive knowledge tests, a traditional multiple-choice test, an implicit-reasoning task, and an instrument requiring participants to indicate whether the various scales were "tests" or "surveys." The nine scales were individually titled and were printed on separate sheets of paper. We administered the package to Air Force recruits and validated the unobtrusive knowledge scales against the participants' ASVAB scores. The ASVAB scores had been collected as part of the standard military recruiting process.

Unobtrusiveness estimates for the scales were obtained by requiring participants to identify each of the nine scales as either a "Test" or a "Survey." The unobtrusiveness data were collected last for the sole purpose of monitoring the "unobtrusiveness" of the Likert scales. Unobtrusiveness data were also collected for the demographic, multiple-choice test, and unobtrusiveness scale (i.e., itself) to provide baseline data to interpret the data obtained for the unobtrusive knowledge scales.

Experimental Materials

The experimental package contained the following nine scales that are listed below and described in their order of administration.

Demographics

The first scale was entitled *Self-Descriptive Information* and required individuals to provide the type of demographic information that is frequently requested in survey questionnaires, e.g., age, gender, ethnicity, automobile ownership, etc.

Unobtrusive Knowledge Tests

The next five instruments consisted of the unobtrusive knowledge tests that were the primary focus of interest for this validation. The five unobtrusive knowledge scales are listed below, with the corresponding content domain that required a rating. Scale length is also reported in terms of the number of items and words used to present the items.

1. *Military Positions*, the size of various Army job families, 15 items and 21 words;
2. *Word Frequency*, the frequency of usage of various English words, 30 items and 30 words;
3. *Excellence*, the connotations of terms implying degrees of excellence, 15 items and 15 words;
4. *Auto Reliability*, the relative reliability of various automobiles, 18 items and 18 words;
5. *Miles per Gallon (MPG)*, the fuel economy of various automobiles, 18 items and 37 words.

The unobtrusive knowledge tests were printed on a scannable paper to simplify data entry. Next to each item was a rectangular box, 5 mm × 122 mm; and

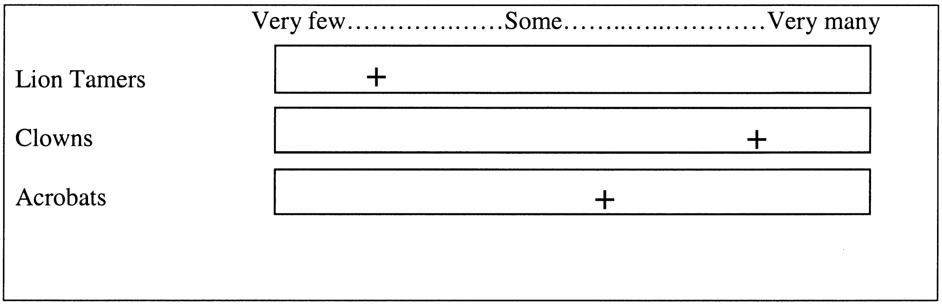
Very few.................Some.......................Very many
Lion Tamers
Clowns
Acrobats
+
+
+

Figure 1. Example of an unobtrusive knowledge scale. Please estimate the number of circus employees you think are in the positions listed below. Indicate your answer by drawing a dark " + " on the spot that provides your best estimate. Use the entire scale because the items span the entire range. In the example, the +'s represents what a respondent might draw.

participants responded by drawing a cross in the box. An example item is provided in Fig. 1. Each rectangular box contained 30 scannable circles that were invisible, adjacent, and 4 mm in diameter. The circles were used to quantify the position of a cross. A cross could be placed anywhere in a box and could correspond to either one or two circles depending on whether it covered multiple circles. A computer program estimated the position of the cross by averaging the value of the circles corresponding to the cross, e.g., a cross covering circles 6 and 7 would correspond to 6.5 "circle units."

Two of the scales, *Word Frequency* and *Excellence*, were intended to coincide with the Verbal factor that is obtained by factoring the ASVAB (cf. Kass, Mitchell, Grafton, & Wing, 1983; Legree, 1995). The other three scales were developed to coincide with the ASVAB Technical factor. (The ASVAB factor structure is described in the following section.) Multiple tests were developed for each factor to allow for some scale failure.

Traditional Multiple-Choice Test

The seventh scale, *Military Knowledge*, was a 33-item multiple-choice test of military knowledge, which required 737 words to present the items. It was constructed by using general military references to develop items.

Unobtrusive Inductive-Reasoning Task

The eighth scale was a series completion task entitled *Guessing Game* that required participants to finish 22 sequences of eight Xs and Os. The task was similar to an opinion survey in that it required participants to indicate how most people would continue the sequence. An example sequence is "XXOOXXOO?" This task can be viewed as an inductive-reasoning task because a response must be implicitly generated by the structure of the sequences as opposed to corresponding to acquired knowledge of how sequences usually appear (Psotka, 1977). Like the Likert knowledge tests, this scale was developed so that it had no computed correct answers or answers that could be retrieved from reference materials. We suspected that this scale would be an unobtrusive measure of inductive-reasoning because it has the unusual characteristic of appearing to be a judgment of others' preferences.

Unobtrusiveness Judgments Task

The final instrument was entitled *Test or Survey* and consisted of a questionnaire that required participants to indicate whether each of nine scales (the previous eight scales and this final scale) had appeared to be either a "test" or a "survey."

Armed Services Vocational Aptitude Battery

The ASVAB is the job classification battery used by the military to assign recruits to military occupations. It is group or individually administered, uses scannable answer sheets, requires approximately 3 h to complete, and is taken by all military recruits. The ASVAB consists of 10 multiple-choice tests named for their content domains: General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto and Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics (EL). All ASVAB tests are power scales except CS and NO, which are speeded.

Four of the ASVAB tests are routinely used to compute the military's operational measure of cognitive aptitude, the Armed Forces Qualification Test (AFQT). AFQT is actually a composite and is calculated by applying unit weights to the following four ASVAB tests: AR, MK, WK, and PC.

Previous factorings of the ASVAB have described three and four first-order factor solutions (cf. Kass et al., 1983; Legree, 1995; Ree & Carretta, 1994). Four-factor solutions provide: a Verbal factor composed of GS, WK, and PC; a Speed factor composed of CS and NO; a Quantitative factor composed of AR and MK; and a Technical factor composed of AS, MC, EL, and GS. (Three-factor solutions merge the Technical and Quantitative factors.) The first-order factors are moderately correlated and load on a single second-order factor, Psychometric *g*. As stated above, the unobtrusive knowledge test domains were selected to overlap with the ASVAB Verbal and Technical factors.

Subjects

Participants were 288 Air Force recruits and were administered the experimental package at Lackland Air Force Base. The recruits had taken the ASVAB during the Air Force enlistment procedure and their scores were contained in Defense Department records.

Procedure

Data were collected over a 2-month period between 7:00 and 9:00 A.M. Prior to participating, individuals read and signed a privacy act statement explaining that their participation was voluntary. The privacy act statement described the scales as "tests" and the recruits were never led to believe that the instruments were "surveys" as opposed to "tests." Subjects were seated in a classroom and were instructed to follow the instructions

Table 1. Internal Consistency Reliability Estimates for the Unobtrusive Tests

Instrument	*Reliability*
Military Positions	.63
Word Frequency	.74
Excellence	.87
Auto Reliability	.50
Miles per Gallon	.86
Implicit-Reasoning	.51

contained in the experimental package. Participant social security numbers were used to obtain ASVAB scores from Defense Department records.

Results

Data Reduction

The procedure used to score the unobtrusive knowledge tests is dissimilar from those used to score most tests. The procedure produces interval data for each item that represents the distance between the subject's rating and a reference value for that item.

However, several data transformations are required to eliminate response bias and to score tests for which answer keys are not already available. If ignored, response bias could have a dramatic effect for subjects who only use part of the rating scale. For example, if the ratings of a particular subject were biased towards the left end of the scale, e.g., range in position from 1 through 15 (in terms of circle units) as opposed to 1 through 30, then the distances calculated for all the responses, except those at the far left end of the scale, would be overestimated.

To resolve the response bias problem, the ratings produced by each subject were transformed within scale to yield z-scores. This transformation resulted in the ratings for each subject for each scale having a mean of 0 and a standard deviation of 1.0. A similar transformation was conducted on the reference ratings that quantify the relative appropriateness of the alternatives. The reference ratings were computed as the mean rating for each alternative across the 288 subjects. Individual differences were computed as the mean item distance between a participant's ratings and the reference values, i.e., absolute value. Better performance is indicated by lower values, and a mean distance of 0 would indicate perfect performance.

The five unobtrusive knowledge tests and the implicit-reasoning task were scored according to the above procedure. The multiple-choice test, *Military Knowledge*, was scored as a proportion correct measure. Neither the *Self-Descriptive Information* (demographic), nor the *Test or Survey* (unobtrusiveness) scales were scored.

Reliability, Correlational, and Descriptive Statistics

Table 1 contains internal consistency reliability estimates for the experimental scales. The *Auto Reliability* scale suffered from low reliability, and the factor analyses were conducted both with and without this scale. In general, the *Auto Reliability* scale had

Table 2. Correlation Matrix[a]

	ASVAB Tests										*Unobtrusive Knowledge Tests*					
	GS	*AR*	*WK*	*PC*	*NO*	*CS*	*AS*	*MK*	*MC*	*EL*	*MP*	*WF*	*EXE*	*REL*	*MPG*	*IRS*
GS	–	72	80	69	52	45	64	69	69	76	45	61	47	03	57	25
AR	40	–	71	67	63	52	53	83	68	66	36	60	43	10	56	31
WK	56	26	–	80	62	55	53	67	59	68	37	65	50	11	56	29
PC	30	25	48	–	61	56	42	64	52	57	26	62	42	08	50	24
NO	−07	19	−09	07	–	70	31	62	41	42	21	44	33	06	35	24
CS	04	20	07	14	52	–	23	52	34	34	22	39	29	14	36	30
AS	51	35	31	22	−12	−02	–	42	74	75	42	40	27	01	48	15
MK	37	55	20	15	29	23	20	–	60	59	28	53	41	02	46	33
MC	54	53	37	33	−05	07	64	39	–	74	45	52	35	03	51	25
EL	57	41	36	21	−07	04	64	29	62	–	39	53	36	08	52	20
MP	36	23	23	10	−06	06	33	11	35	26	–	31	16	08	35	13
WF	32	30	34	30	−02	04	18	16	33	25	18	–	43	03	47	22
EXE	24	17	25	13	−01	03	09	13	15	11	03	23	–	09	35	14
REL	−04	08	05	02	00	10	−02	−05	01	05	07	−02	05	–	05	06
MPG	34	33	29	22	−05	09	34	16	34	30	24	24	16	01	–	18
IRS	09	18	05	07	09	20	06	21	16	07	06	09	02	05	06	–

Note: [a]Sample correlations are in the lower triangle and population estimates are in the upper triangle. Decimal points are omitted, and correlations above .17 are significant at the .05 level. The correlations are reflected so that superior performance on any test would always be positively correlated with superior performance on any other scale in the test battery.

Table 3. Correlation Matrix of Latent Variables

	Sample Correlations			*Population Estimates*	
Scales	*AFQT*	*ASVAB-g*	*UKT-g*	*AFQT*	*ASVAB-g*
Military Positions	.26	.33	.53	.43	.58
Word Frequency	.39	.37	.76	.60	.63
Excellence	.24	.22	.45	.41	.42
Reliability	.04	.04	.07	.07	.07
Miles Per Gallon	.37	.39	.69	.57	.66
Implicit-Reasoning Scale	.19	.21	.19	.32	.40
UKT-g	.52	.54	–	.74	.80

very low coefficients associated with it. We considered reporting analyses without this scale, but decided to include it to provide a more comprehensive coverage of our findings. As a general statement, the decision to include or exclude the task had no impact on our general conclusions.

Table 2 reports the correlations among the unobtrusive scales and the ASVAB tests, and Table 3 contains correlations between the unobtrusive scales and the ASVAB composites, which are described below. Unlike most cognitive tests, low scores indicate better performance on the unobtrusive tests. To simplify our presentation, we reflected the correlations involving the unobtrusive scales so that superior performance on any test in the battery would always be positively correlated with superior performance on any other test in the entire battery of 16 tests. This transformation allows the reader to more easily locate any exceptions to positive manifold.

The multivariate correction for range restriction was used to estimate the population correlations between the experimental scales and the ASVAB tests because range restriction due to military entrance requirements substantially attenuates all ASVAB correlations (Johnson & Ree, 1994). The bivariate correction for range restriction was used to estimate population estimates for correlations involving ASVAB composites. The population correlation estimates are reported in the upper triangular portion of the matrix presented in Table 2 and in the last two columns of Table 3.

Tables 2 and 3 indicate that four of the unobtrusive knowledge scales correlated substantially with the ASVAB tests and composites. Sample correlations ranged to .39 and population correlation estimates ranged to .65. The only unobtrusive scale that did not correlate substantially with the ASVAB tests or composites was the *Auto Reliability* scale, which suffered from low reliability.

Table 4 reports the percentage of participants who viewed each scale as a "test" as opposed to a "survey." These data were collected using the final *Test or Survey* scale after all the other scales had been completed. The percentages corresponding to the *Self-Descriptive* scale, the *Test or Survey* scale, and the *Military Knowledge* (multiple-choice) test are reported to provide baseline values with which to interpret the data for the unobtrusive scales.

The principal finding in Table 4 is that between 13 percent and 22 percent of the participants described the unobtrusive knowledge scales (Word Frequency, Excellence, MPG, and Auto Reliability) as "tests." These percentages are meaningful because they are in the range of values that correspond to the percentage of individuals who describe the

Table 4. Percentage of Air Force Recruits Viewing Unobtrusive Tests as "Tests" as Opposed to "Surveys"

Scales	*Percentage*
Unobtrusive scales	
Military Positions	60
Word Frequency	21
Excellence	22
Auto Reliability	13
Miles Per Gallon	13
Implicit-Reasoning	51
Conventional knowledge test[a]	
Knowledge of Military	92
Conventional survey scales[a]	
Self-Descriptive Information	10
Test or Survey	25

Note: [a]Entries included as baseline values to interpret the experimental scale data.

Table 5. Goodness-of-Fit Statistics for the Four-Factor Model

Statistic	*Value*
Chi-square statistic	179.03
Degrees of freedom	98
Probability	.000001
Goodness-of-Fit index	.92
Root mean square residual	.043

Self-Descriptive and *Test or Survey* scales as "tests" (10% to 25%). The one exception to this finding was the *Military Position* scale (60%). The data also indicate that a moderate percentage of participants viewed the inductive-reasoning task as a "test" (51%). As expected, most participants (90%) described the multiple-choice test as a "test." The data in Table 4 validate the assumption that the Likert scales act as unobtrusive knowledge tests because most of the coefficients fall in the range obtained for the *Self-Descriptive* and *Test or Survey* scales.

Confirmatory Factor Analyses (CFA)

Lisrel was used to analyze the factor structure of the experimental battery in relation to the ASVAB (Jöreskog & Sörbom, 1993). We hypothesized that the unobtrusive scales would load on factors defined by the ASVAB tests because the content domains corresponding to the unobtrusive scales were chosen to overlap with ASVAB content domains. Alternate hypotheses included the possibilities that the unobtrusive knowledge scales would either correspond to a separate factor reflecting their unique Likert method or would load in some unexpected fashion.

The factor structure of the experimental battery was assessed by including the experimental and the ASVAB tests in a CFA. The *Word Frequency*, *Excellence*, and *Inductive-Reasoning* scales were hypothesized to load on a Verbal factor while the *MPG*, *Military Positions*, and *Automobile Reliability* scales were expected to load on the Technical factor. We used Lisrel modification information to delete trivial paths and to create additional links.

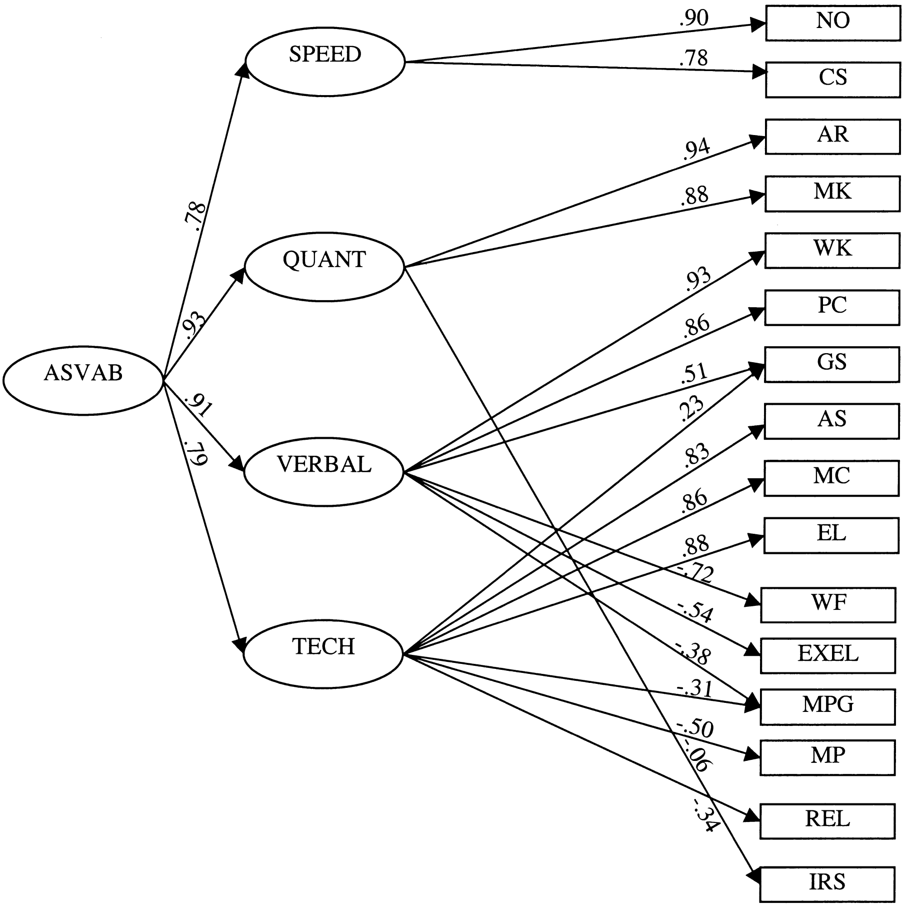

Figure 2. Structure of the four-factor model.

We conducted the CFA twice, once using the sample correlation matrix, and once using the population matrix; similar conceptual results were obtained for both factor analyses. The main difference between the two factor analyses (sample vs. population correlations) was that lower loadings were obtained for the sample matrix, as would be expected due to the range restriction. Another minor difference was that the sample correlations for one of the speeded ASVAB tests, CS, was problematic from a modeling perspective; however, this finding had little impact on our general conclusions because our hypotheses did not relate to the ASVAB speeded factor. In the interest of economy, we are reporting only the analyses for the population matrix.

The "best" model that could be developed was very similar to the hypothesized model except that: (1) the *Inductive-Reasoning* task loaded on the Quantitative factor as opposed to the Verbal factor, and (2) *MPG* loaded on both the Technical and Verbal factors. Table 5 contains the Goodness-of-Fit statistics calculated for this model and Fig. 2

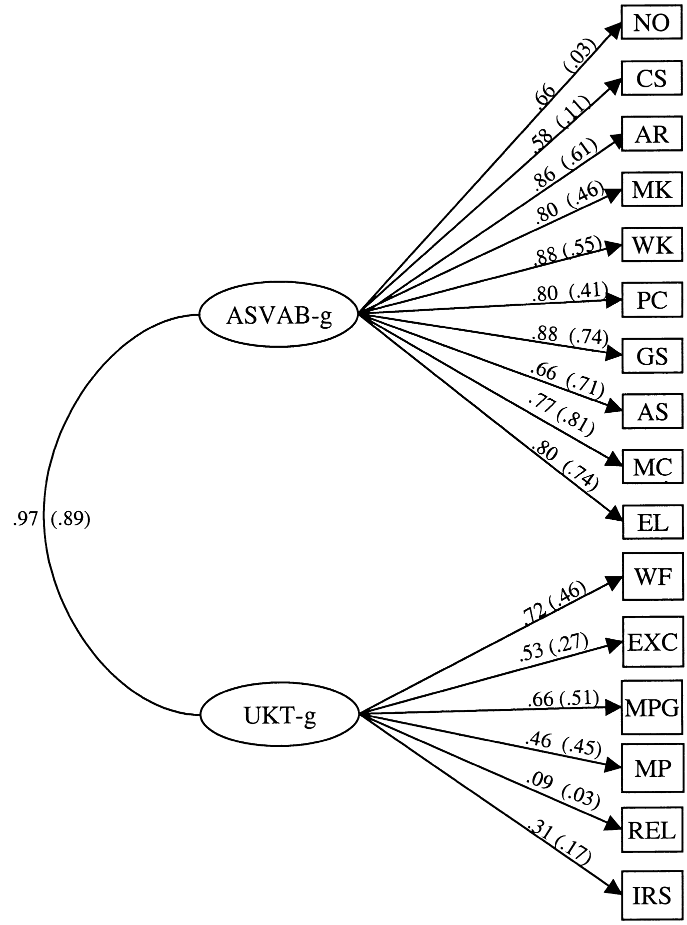

Figure 3. Structure of the two factor model. Path coefficients based on the sample correlation matrix are reported in parentheses.

contains the corresponding factor structure. The factor structure generally confirms our expectation that the unobtrusive scales would load on factors that are defined by conceptually related ASVAB tests.

Equivalence of Latent Variables (Psychometric *g*'s) Defined by the Two Batteries

While the above analyses demonstrate that the unobtrusive scales measure individual differences in knowledge domains that correspond to the ASVAB factor structure, they do not estimate the correlation between estimates of Psychometric *g* that would be obtained by independently factoring the two batteries. To address this question, we specified a very simple model that contained two latent variables: the first corresponding to Psychometric *g* as defined by the ASVAB tests and the second to Psychometric *g* as defined by the unobtrusive scales. We then estimated the path between the two latent

Table 6. Goodness-of-Fit Statistics for the Two-Factor Model

Statistic	*Value*
Chi-square statistic	385.32
Degrees of freedom	103
Probability	.0000004
Goodness-of-Fit index	.83
Root mean square residual	.086

variables. This analysis was conducted twice by using the population and sample correlation matrices.

The results are reported in Fig. 3 and Table 6. The path between the two latent variables estimates the correlation between these two Psychometric *g*s. It is important to understand that this path represents the theoretical correlation between the two latent variables *if they were measured without error*. The path coefficient calculated for the population matrix, .97, demonstrate a high degree of convergence between the two approaches to estimating Psychometric *g* (A slightly lower coefficient, .89, was obtained by analyzing the sample correlation matrix). It follows that the two approaches measure essentially the same higher-order factor, which is usually referred to as Psychometric *g*.

Convergence of the Two Observed Psychometric *g*'s from the ASVAB and Unobtrusive Scales

While the preceding analyses demonstrate the theoretical convergence (i.e., given error-free measurement) between the Psychometric *g*s defined by the ASVAB and Unobtrusive Scale Batteries, they do not estimate the level of convergence between observed estimates of Psychometric *g* that would be obtained by independently using the two test batteries.

We addressed this issue by using SPSS to factor the Unobtrusive Scale Battery and to compute the corresponding factor scores. We also factored the 1980 ASVAB norming data (U.S. Department of Labor & Bureau of Labor Statistics, 1997) to calculate the population loadings of the ASVAB tests on Psychometric *g*. The procedure and factor loadings are described in Appendix A. These factor loadings were then used to compute composite scores for our sample that would correspond to the Psychometric *g* factor scores computed using the ASVAB population norming data. This procedure was followed to obtain ASVAB-based Psychometric *g* composite scores for which the population variance would be known.

These composites were named ASVAB-*g* and UKT-*g* and correlations involving these composites are reported in Table 3. For the present purpose, the most important correlation in Table 3 is between these two composites, and it is equal to .54, $p < .001$. The sample and population standard deviations were equal to 1.0 and .49, and the bivariate correction for range restriction was used to estimate the population correlation between ASVAB-*g* and UKT-*g*, .80.

Regression Analyses

The above analyses show that the Unobtrusive Battery can be used to provide an accurate estimate of Psychometric *g*. However, we expect that in many survey applications, only a subset of the battery could be administered due to space and time limitations. We therefore

Table 7. Summary of the Regression Analyses

Multiple R	*Adjusted R*	*Number of Variables*	*Method to Enter*	*Unobtrusiveness Range*	*Variables*
Equations using UKT-g as the dependent variable					
.92	.91	2	Stepwise	21–13	WF, MPG
.97	.96	3	Stepwise	21–60	WF, MPG, MP
.99	.99	4	Stepwise	21–22	WF, MPG, MP, EXC
.95	.94	3	Enter	21–22	WF, MPG, EXC
.95	.95	4	Enter	21–51	WF, MPG, EXC, IRS
Equations using ASVAB-g as the dependent variable					
.48	.48	2	Stepwise	13–21	MPG, WF
.53	.52	3	Stepwise	13–60	MPG, WF, MP
.55	.54	4	Stepwise	13–51	MPG, WF, MP, IRS
.48	.48	3	Enter	13–22	MPG, WF, EXC
.52	.51	4	Enter	13–22	MPG, WF, IRS, EXC
Equations using AFQT as the dependent variable					
.48	.47	2	Stepwise	21–13	WF, MPG
.50	.49	3	Stepwise	21–51	WF, MPG, IRS
.51	.50	4	Stepwise	21–51	WF, MPG, MP, IRS
.51	.50	4	Enter	21–22	WF, MPG, IRS, EXC

used a stepwise regression procedure to identify optimal subsets of the unobtrusive scales consisting of between two and four tests. We also identified and included alternate subsets that might be preferable because they minimize the obtrusiveness ratings of the scales. The results are summarized in Table 7 and show that reasonably accurate estimates of cognitive aptitude could be obtained with between two and three unobtrusive scales.

Discussion

We were stunned by the overwhelming success of the unobtrusive knowledge tests in accurately estimating Psychometric *g*! Our goal had been to provide a procedure that could be used to estimate general cognitive aptitude for survey research purposes, and the regression analyses show that our success was largely due to the Word Frequency and MPG scales. Because survey data are not generally used to make decisions affecting individuals, we were prepared to accept as useful a procedure associated with a population validity estimate in the .5 to .6 range. We had felt that even relatively inaccurate cognitive aptitude test data could provide useful information for the development of structural equation models to understand better social and market phenomena.

However, the population validity estimate between the cognitive aptitude scores based on the unobtrusive scale battery and the ASVAB is equal to .80. This value provides an extremely strong endorsement for using unobtrusive scales to measure general cognitive aptitude because .80 falls in the range of correlations typically obtained when high quality cognitive aptitude batteries are administered to the same population. In fact, using tests that have validity coefficients in this range to support decisions affecting individuals is easily justified. Our results demonstrate that the unobtrusive test battery could be used to accurately measure individual differences in general cognitive aptitude for survey research purposes. It follows that this information could help analysts to better understand a variety of market and social phenomena.

The CFA, which is summarized in Fig. 2 and Table 5, provides a more technically oriented endorsement of the approach used to develop the unobtrusive knowledge tests because the experimental scales loaded on the ASVAB factors in accordance with their content domains. This finding demonstrates the general validity of the procedure and indicates that the method could be used to develop scales for additional domains where a conventional test might be inappropriate due to the obtrusive nature of most conventional tests.[2]

Because the unobtrusive knowledge tests were developed to correspond to multiple ASVAB factors, it follows that factoring the unobtrusive test battery should have provided a higher-order factor that would be very similar to one obtained by factoring the ASVAB. The second factor analysis addressed this question and the results show an extremely strong relationship between the two sets of factor scores corresponding to the unobtrusive test battery and the ASVAB. These results are summarized in Fig. 3 with the path between ASVAB-*g* and UKT-*g* being equal to .97.

The regression analyses are important from a practical standpoint because they show that a very accurate measure of Psychometric *g* could be obtained with between two and three unobtrusive knowledge tests. Because these are highly efficient scales, it is reasonable to expect that an abridged yet accurate test battery could be included in many survey questionnaires. The regression analyses were conducted using a stepwise procedure to empirically identify the most valid sets of scales.

We repeated the regression analyses and replaced the military-oriented knowledge test, *Military Positions*, with alternate scales including the *Excellence* test and the *Unobtrusive Inductive-Reasoning* task. Our results show that these changes resulted in only a slight decrease in validity. Because the unobtrusiveness ratings were much lower for the alternate scales (refer to Table 7), we feel that the alternate sets would be more appropriate for most applications.

It is important to appreciate that most subjects did not characterize the unobtrusive knowledge scales as "tests," instead, they described the scales as "surveys." In fact, the percentages of participants describing the unobtrusive knowledge scales as "tests" compare favorably with the percentages describing the *Self-Descriptive Information* and *Test or Survey* scales as "tests." These data suggest that relatively few individuals would view the unobtrusive knowledge scales as "tests" if they were included in mail-, Internet-, or paper-based surveys. It seems reasonable to assume that even fewer participants would attempt to provide "good" responses by consulting standard reference sources. Since published reference sources are not available for these domains, the possibility that cheating would invalidate the cognitive aptitude data is virtually eliminated.[3]

It is curious to note that a moderate number of the participants, 60 percent, viewed the *Military Positions* scale as a "test." This is a somewhat unique scale in that the participants had recently enlisted in the Air Force. While the scale dealt with Army jobs, it seems reasonable that the scale's military-knowledge domain and the participants' status as new recruits could interact and lead to the expectation that any question regarding the military was a "test."

The number of individuals who described the *Self-Descriptive* and *Test or Survey* scales as "tests," 10 percent and 25 percent, respectively, were slightly higher than expected in that we had anticipated that these percentages would approach 0. Upon reflection, it now seems possible that our background in psychometrics may have acted to over-state the distinction we (the authors) see between "tests" and "surveys." In other words, we were led to consider the notion that non-psychometricians are less likely to distinguish between these terms. In support of this notion, we observed that the term, "test," is used in the guide, *Standards for Educational and Psychological Testing*, which is published jointly by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1985), to loosely describe personality inventories, behavioral observations, and projective techniques, as well as aptitude and achievement scales.

While few survey efforts have incorporated a measure of general cognitive aptitude, such data could provide new insights into relationships among cognition, and a number of environmental and social factors. Many examples of these kinds of analyses can be found in *The Bell Curve* (Herrnstein & Murray, 1994). However, the aptitude data analyzed by Herrnstein and Murray were collected primarily to compute norms for a cognitive test battery and were included in the 1980 National Longitudinal Survey of Youth (NLSY) to help assure that the norming sample would be representative of the U.S. youth population. Thus, the links among the NLSY social and cognitive aptitude variables were not explicitly designed to provide an empirical database to explore cognitive aptitude and its relationship to social and economic phenomena. Well-designed empirical studies of these relationships might alter or modify some of the basic findings.

Regardless of one's perspective on the quality and implications of the Herrnstein and Murray analyses, their results show the importance of relating cognitive aptitude to more typical survey measures, and the need for better experimental control of these measures. Many survey databases exist that could be used to support, deny, or extend these findings if those surveys had collected cognitive aptitude data.[4] The principal scientific merit of these consensual, unobtrusive scales developed in this article, is to explore the vast space of relationships among cognitive aptitude and salient social behaviors that are often the focus of many surveys. From this perspective, the unobtrusive knowledge tests are relevant to advancing the horizontal aspect of research on mental aptitude, i.e., its broad social and economic ramifications. Researchers interested in utilizing the unobtrusive knowledge tests should contact Peter J. Legree at legree@ari.army.mil or petelegree@aol.com.

Appendix A. Loadings Used to Calculate ASVAB-*g* Scores

We obtained the NLSY80 database from the Ohio State University and based our analyses on ASVAB norming data collected for participants who were born before 1962.

Standardized Tests	*Beta Weight*
General Science	.009305
Arithmetic Reasoning	.016458
Word Knowledge	.022749
Paragraph Comprehension	.007214
Numerical Operations	.014469
Coding Speed	.009906
Auto Shop Information	.008960
Math Knowledge	.016958
Mechanical Comprehension	.008992
Electrical Information	.008392

Our analyses utilized the population sampling weights provided with the database. The SPSS principal axis factoring routine was used to extract and rotate four oblique first-order factors. The first-order factors were then entered into a principal component analysis to extract a single higher-order factor. This procedure reflects guidance found in Jensen and Weng (1994). The population mean and standard deviation for this variable are 0 and 1.0.

Notes

1. The views, opinions, and/or findings contained in this article are solely those of the authors and should not be construed as an official Department of the Army or DOD position, policy, or decision, unless so designated by other documentation.

2. Another possible application for this technology is the development of knowledge tests for legally or emotionally charged domains for which individuals may not want to self-disclose, e.g., sexual behavior or drug use. This concept reflects the notion that the presence or lack of knowledge indicates involvement in certain activities. Sometimes, the lack of knowledge is a "good thing"; recently, one of the authors learned that none of the friends of his teenage son could estimate the cost of drugs.

3. Unobtrusive tests might also be useful to evaluate Steele's Stereotype Threat hypothesis, which attempts to reconcile group differences in cognitive aptitude as due to anxiety in manifesting racial stereotypes (Steele & Aronson, 1995).

4. Another reason to disseminate our findings is that databases may already exist that contain "attitudinal" items that could be scored as knowledge questions and may load on Psychometric *g*. It is important to appreciate this possible use of survey information so that appropriate decisions can be made regarding its use.

While we believe the use of this procedure does not rely on the failure of respondents to appreciate that they are taking a test, respondents are nevertheless generally responding to these items as something other then test items. Thus, we caution our readers that we do not recommend that this procedure be used for making personnel decisions. Such a use could only be justified if the respondents had the appropriate understanding that they were, in fact, providing responses that would be used for such a purpose. However, for the purposes we have identified for collecting survey data that is provided anonymously, we believe the use of this procedure is justified.

References

American Educational Research Association (AERA), American Psychological Association, (APA) & National Council on Measurement in Education(NCME). (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Herrnstein, R. J., & Murray, C. A. (1994). *The Bell curve: Intelligence and class structure in American life*. New York, NY: Free Press.

Jensen, A. R., & Weng, L. J. (1994). What is a good *g*? *Intelligence, 18*, 231–258.

Johnson, J. T., & Ree, M. J. (1994). RANGEJ: A Pascal program to compute the multivariate correction for range restriction. *Educational and Psychological Measurement, 54*, 693–695.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kass, R. A., Mitchell, K. J., Grafton, F. C., & Wing, H. (1983). Factorial validity of the Armed Services Vocational Aptitude Battery, Forms 8, 9 and 10: 1981 Army applicant sample. *Education and Psychological Measurement, 43*, 1077–1087.

Legree, P. J. (1995). Evidence for an oblique social intelligence factor established with a Likert based testing procedure. *Intelligence, 21*, 247–266.

Legree, P. J., Fischl, M. A., Gade, P. A., & Wilson, M. J. (1998). Testing word knowledge by telephone to estimate general cognitive aptitude using an adaptive test. *Intelligence, 26*, 91–98.

Psotka, J. (1977). Syntely: Paradigm for an inductive psychology of memory, perception, and thinking. *Memory and Cognition, 5*, 553–560.

Ree, M. J., & Carretta, T. (1994). Factor analysis of the ASVAB: Confirming a Vernon-like structure. *Educational and Psychological Measurement, 54*, 459 – 463.

Steele, C., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 781–797.

U.S. Department of Labor & Bureau of Labor Statistics. (1997). *The national longitudinal surveys*. Columbus, OH: Center for Human Resource Research, The Ohio State University.