Journal of Applied Psychology
2000, Vol. 85, No. 6, 940–955

0021-9010/00/$5.00 DOI: 10.1037//0021-9010.85.6.940

Validity of Measures of Cognitive Processes and General Ability for Learning and Performance on Highly Complex Computerized Tutors: Is the *g* Factor of Intelligence Even More General?

Mary Roznowski and David N. Dickter
The Ohio State University

Sehee Hong
University of California, Santa Barbara

Linda L. Sawin and Valerie J. Shute
Brooks Air Force Base

Theoretical arguments and analyses from 2 studies provide compelling evidence that computerized measures of information-processing skills and abilities are highly useful supplements to more traditional, paper-based measures of general mental ability for predicting individuals' capacity to learn from and perform on highly challenging, multifaceted tutors. These tutors were designed to emulate learning and performance in complex, real-world settings. Hierarchical confirmatory factor analysis provided evidence that a general, higher order factor model with general ability at the apex could quite adequately and singularly account for the individual-differences data, both traditional and cognitive-process measures. Results are interpreted in light of the utility and generality of human cognitive abilities.

The measurement of human cognitive abilities has evolved dramatically in the past century of research on individual differences, yet in some ways it has remained remarkably stable. Many of our present day measures appear highly similar to the first measures introduced during Alfred Binet's and J. McKeen Cattell's day, for instance. However, part of this stability may merely be an accidental byproduct of technological limitations such as a reliance on paper-based measures. Another cause of, or at least a contributor to, the consistency has been psychometricians' close adherence to factor analytic frameworks and methodologies to guide test construction. The coupling of these considerations has perhaps resulted in somewhat limited movement and may have curtailed advancement in the measurement of critical latent traits such as human intellectual ability and other cognitive attributes. However, in recent years, owing both to rapid advancements in computer-assessment technologies and to developments in cognitive psychology and other disciplines, novel measures have been introduced that may serve to promote movement in the testing of important individual differences (see Mead & Drasgow, 1993).

Another critical strain of thinking and research is relevant here. Recent work by Schmidt and Hunter (1998) summarizing 85 years of personnel selection research indicates that although many individual qualities and attributes are important, general intellectual ability remains the single best predictor of job and training performance. The combination of measures of general ability and other assessments or selection procedures yields mean validities in the impressive range of .63–.65, most of which is attributable to variance due to general ability differences among candidates. These authors also reiterated the fact that even small incremental validity gains translate directly into utility benefits for organizations.

Indeed, we have considerable cumulative evidence of the utility of measures of general ability for predicting both training and job performance. Furthermore, Schmidt and Hunter (1998) pointed to the impressive conceptual and historical foundations forged around the general ability construct. A critical question becomes obvious, however: How well do current measures, although effective in a predictive sense, actually reflect or mirror the construct space identified in this vast body of conceptual thinking? It becomes useful, then, to touch briefly on aspects of this theoretical foundation for general ability as it relates to the purposes of the present study and the vital question of construct coverage. The current article attempts to answer the question implied here. That is, is there additional construct variance that might be culled to improve the measures commonly used in the field as well as the predictability of critical criterion space?

We realize that we are at risk of committing reification errors and reopening long-standing debates concerning the reality status of constructs (Allport, 1937, 1966; Meehl, 1986). However, a discussion of the conceptual nature of the general factor of intelligence and how well and extensively the field's current measures reflect that hypothetical factor is critical. Almost a century ago, Spearman (1904) stated that a primary "part of intelligence is

Mary Roznowski and David N. Dickter, Department of Psychology, The Ohio State University; Sehee Hong, Department of Education, University of California, Santa Barbara; Linda L. Sawin and Valerie J. Shute, Armstrong Laboratory, Brooks Air Force Base, Texas.

David N. Dickter is now at AT&T, Basking Ridge, New Jersey. Linda L. Sawin is now at Boeing, Concord, California. Valerie J. Shute is now at knowledgeplanet.com, Puget Sound, Washington.

Correspondence concerning this article should be addressed to Mary Roznowski, 372 Brynhild Road, Columbus, Ohio 43202. Electronic mail may be sent to roznowski.2@osu.edu.

played by [one's] ability to handle, not merely abstract ideas, but above all symbols" (p. 211). In another early definition of general mental ability, Pintner (1921) included the "ability of the individual to adapt . . . adequately to relatively new situations in life" (p. 129). As we discuss later, both of these notions, symbolic representation and adaptability, are critical to the conceptual nature of general intelligence and are central to the measures tested herein. Furthermore, and also critical, one of the primary elemental processes of intelligence Spearman (1923) elucidated is the education of correlates—that is, the attributes of stimuli that are seen as identical to or similar to each other or are compared or related in some way (see Jensen, 1998).

In his early, classic writing, Spearman (1904) stated that all tests of an intellectual nature index a single underlying factor and do so to varying degrees. This seemingly simple notion is critical and forms the basis for the current research. Spearman (1904, 1927) was also first to point out that tests have differing levels of intercorrelation with other tests. An ordering of variables on the basis of their levels of intercorrelation can be said to signify that each variable can be accounted for, to some quantifiable extent, by a single factor common to all the variables (Spearman, 1904). This factor is the ubiquitous *g*, and, in Spearman's (1904) words, tests have different levels of saturation with this *g* factor. Indeed, Spearman (1927) indicated that certain cognitive tasks better assess or reflect *g* than others do, and those that best measure the hierarchical factor indexing *g* are tests that require heavy use of reasoning and language understanding and an ability to deal with quantitative relationships.

Relevant to the current thread from both a measurement and a theoretical perspective, Humphreys (1971) has long discussed the need for breadth and heterogeneity in the measurement of intelligence. This is necessary in part to spread out unwanted bias factors or contaminants but also to increase construct-relevant variance in measures and batteries. Such breadth is needed to build up measurement of the factor space of interest, which in the case of general ability is believed to be quite broad, diverse, and very rich (see also Jensen, 1998).

Along these lines, Cronbach (1989) stated that current tests may perhaps be overly academic, neglecting problem solving and the gamut of potential intellective activities that are available. He pointed to the need, also articulated by Humphreys (1984), for assessing widely divergent kinds of activities, all of an intellectual nature. Indeed, Humphreys and Parsons (1979) have indicated that tests could indeed be more broadly constructed than they currently are. For instance, Humphreys and Parsons have shown convincingly how traditional Piagetian tasks (conservation of substances, rotation of squares, etc.) might be used to improve current ability measures designed for youthful populations. Cronbach (1989) has also raised a very compelling question: "Has the historical process located [the] center [of the intelligence construct] where it should be?" (p. 151). He has indicated that for the test to be an uncontaminated estimate of the underlying trait, a measure must point to the locus in the construct space that matches a proposed conceptual foundation. This, of course, is the classic construct validity problem, as defined and discussed by Campbell and Fiske (1959; Fiske & Campbell, 1992), Cronbach (1989; Meehl & Cronbach, 1955), Messick (1995), and many others.

Notwithstanding Boring's (1923, 1950) early reductionistic view that intelligence is simply and only what intelligence tests measure, it is important to begin to revisit the question aimed at determining whether general intelligence is indeed not much more than what our current tests measure. To wit, Humphreys (1971) has stated that intelligence is best viewed as the entire repertoire of intellectual information, knowledge, and skills available to a person at a particular point in time. Repertoire is indeed a very apt metaphor for the broad construct space of intelligence. Furthermore, all tests of an intellectual or cognitive nature sample this repertoire and do so to varying degrees. Humphreys (1985) has also indicated that some behavioral acts are simply more central to the repertoire than others are. For example, knowing word meanings is likely more central to the repertoire than is rote memory of nonsense syllables or recognition of figural content. Thus, one's goal in measuring intelligence is best met by systematically sampling this repertoire and by doing so with tests that attempt to sample as large and diverse a segment of the construct space as possible. Intelligence is also said to affect the development of many different types of skills and knowledge. Thus, the more such different types of skills and knowledge a measure taps, the more complete and construct valid the measure is (see Brody, 1992; Jensen, 1998).

It is critical, then, to ask whether important elements of human cognitive abilities have been "left out" because of various technological limitations, tradition, or other factors. For instance, were Lubinski and Dawis (1992) correct in stating that there may simply be more to general ability than typically surfaces in traditional measures such as the Armed Services Vocational Aptitude Battery (ASVAB; see Dunnette, 1976)? Also relevant is the fact that Embretson (1983) has stated that test design should be carried out by "specifying the aspects of individual differences that a test measures by constructing and/or selecting items according to their substantive properties" (p. 3). One way to achieve these various goals is to measure in a highly systematic way the array of cognitive elements and processes that theoretically might reflect, and thus are reflectors of, cognitive ability. Accordingly, for several years, a major research effort involved in developing and evaluating novel and efficient ways of measuring human cognitive abilities and skills has been the Air Force's Learning Abilities Measurement Program (LAMP; see Kyllonen et al., 1990). LAMP researchers made extensive use of computer administration of a unique collection of cognitive tasks to measure a broad range of cognitive skills and abilities, including information-processing speed and working-memory capacity. From its inception, the program attempted to maintain the dual goals of modeling cognitive learning and developing improved tests for selection and classification (Kyllonen & Christal, 1989).[1]

The primary purpose of the current study is to assess the tenability of building highly trait-relevant breadth into measures, thereby better capturing or reflecting the relevant factor space of human mental abilities (see also Lubinski & Dawis, 1992). We

[1] During the time period this research was carried out, the LAMP research group included Patrick Kyllonen, Raymond Christal, Scott Chaiken, Lisa Gawlick, Carmen Pena, Valerie Shute, Bill Tirre, and Dan Woltz, as well as the programming staff: Henry Clark, Trace Cribbs, Rich Walker, Cindi Garcia, Jo Ann Hall, Janice Hereford, Terri Purdue, and Karen Raouf. Our sincere appreciation and regard goes to each of these individuals.

attempt to do this by evaluating measures developed as part of LAMP's efforts, as well as by examining the predictive validities of scores from a subset of measures from LAMP's Cognitive Abilities Measurement (CAM) test battery (Kyllonen et al., 1990) and comparing the scores with traditional, paper-based measures of general ability (*g*). Thus, part of our overall goal is to understand individual differences in cognitive ability better by assessing a broad range of information-processing components. Indeed, researchers are now able, using microcomputer technology, to study abilities that were simply not amenable to research in prior times using traditional assessments. Furthermore, we make an attempt at determining whether *g* is indeed configured more broadly than traditional conceptualizations and current operationalizations indicate. The battery of diverse, novel cognitive-process measures analyzed here enables us to begin to address this critically important measurement, psychometric, and theoretical issue.

As part of this work, we needed criteria that would allow important predictive and construct validity questions to be answered. Because of the dependence in modern organizations on measures of individual cognitive attributes for selection and placement, as well as the utility of general ability measures, we chose the context of technical performance as a criterion domain in which to evaluate the cognitive-process measures. We sought criteria that emulated as much as possible the technical performance and learning environments present in modern organizations. Such criteria both allowed for assessment of the utility of the cognitive-process measures for applied purposes and served as an aid in answering several construct-validation questions regarding both traditional and experimental measures. Criteria meeting these necessities were performance scores from two highly complex, computerized tutors that assess learning of electricity concepts (Study 1; see Shute, 1993a) and flight engineering knowledge and skills (Study 2; see Shute, 1993b). Because training and work activities in today's organizations have both begun to move increasingly toward information technology and computer modalities, and given the increased information-processing demands of jobs, we believe that such diverse and individualized tutors reflect the training and performance environments found in many job settings. They also enable researchers to measure with considerable experimental precision the acquisition of knowledge, skills, and abilities that are relevant to many real-world work situations (Campbell, 1990; Christal, 1991).

Mastery on the tutors was part of the criterion definition for our study and was defined as examinees correctly solving several consecutive problems on a given principle or module. Thus, participants could not complete the study without actually reaching a criterion of proficiency on the studied material. After they completed the tutors, participants were given a comprehensive test on the principles that they had learned, and their overall accuracy was recorded. The total time it took individuals to complete the electricity concepts tutor (Study 1) was also recorded. The final accuracy tests contained both true–false and multiple-choice items. Content of both qualitative and quantitative natures was also assessed. Tutor scores were thus believed to reflect individual differences in learning ability and understanding of novel knowledge domains. They were clearly emulations of the everyday demands of many workplace environments.

The LAMP researchers used a detailed taxonomy for designing and selecting tests for the cognitive-process battery. Kyllonen (1991) compiled a very useful description of the basic principles that are commonly followed in creating computerized test batteries. The LAMP researchers used an information-processing perspective rather than a traditional factor analytic framework to guide test development, item writing, and test selection. In the work described here, a "radex" model guided the theory-based test development; this allowed novel connections between the experimental and differential literatures (Snow, Kyllonen, & Marshalek, 1984). A taxonomy (Kyllonen, 1994) resulting from this two-dimensional radex modeling of abilities contained six information-processing components, each with three stimulus or content domains. This resulted in 18 potential test types. The six cognitive processing factors assessed were working memory (WM), general knowledge (GK), processing speed (PS), fact learning (FL), skill learning (SL), and induction (IN). Three classic content–stimulus domains were used for the tests: verbal, quantitative, and spatial. These content domains were fully crossed with the six processing factors, yielding the 18 test types, which included, for instance, skill learning–verbal and skill learning–quantitative. As Kyllonen (1991) has pointed out, traditional taxonomies and measures do not effectively and completely cross content factors with process factors, a fact that was accomplished with the CAM battery. For instance, verbal and quantitative induction (both assessed here) are rarely present in traditional measures. All potential test types with the exception of two were examined in our research. Similarly, Lubinski and Dawis (1992) pointed out that tests such as the formerly widely used General Aptitude Test Battery (GATB; see Dunnette, 1976) or the ASVAB do not uniformly or adequately measure all regions of the radex dimensional space that reflects the purported organization of human abilities. Indeed, Humphreys (1979) has stated compellingly that the wide variety of human behaviors that can be labeled as cognitive or intellectual should be sampled, if possible, in the tests used by the field. It indeed may be true that the construct representation scope of the field's measures have been limited, in part because of technological and methodological limitations as mentioned previously.

In addition to representing cells in this taxonomy, the tests in our battery correspond closely to current paradigms in the experimental psychology and cognitive-science literatures. For instance, long-term episodic memory (Loftus, 1983; tested here with fact learning), working memory (Anderson, 1983; Baddeley, 1986; Woltz, 1988), skill learning and acquisition (Anderson, 1983), and induction and inference learning (Holland, Holyoak, Nisbett, & Thagard, 1986) are measured by the various tests. Item "faceting" is incorporated, such that aspects of difficulty and complexity are varied in a systematic fashion within tests (see Irvine, Anderson, & Dann, 1990). For instance, when one incorporates negation, complexity and item–trial difficulty levels and test error rates are increased in a predictable way, whereas proportion correct is typically decreased. (An example of negation would be "X is not a member of Set A or Set B.") Brief descriptions of the six cognitive component–process types and the three stimulus domains follow.

General Knowledge

General knowledge refers to a set of measures of the common information knowledge accessible to individuals. This domain is intended to assess the depth and breadth of individuals' declarative

knowledge base and their ability to answer general knowledge questions. Example domains include information regarding measurements and U.S. geography.

Processing Speed

This domain assesses the speed it takes individuals to retrieve, transform, and respond to stimuli. In our battery, participants decided whether two presented stimulus words conformed to a sequence specified in a sentence at the top of the computer display. Faceting and difficulty were incorporated into items with the use of minus signs indicating, for instance, that a reversal of a presented ordering was required.

Working Memory

This domain assesses temporary storage of information being processed during a cognitive task. Working memory is believed to include a general representation of a given situation "without necessarily containing an exact record of all details of the information that has been presented" (Hunt & Pellegrino, 1985, pp. 220–221). Such measures should require "simultaneous processing and storage of information" (Baddeley, 1986, pp. 34–35). Working-memory capacity has been shown to be a good predictor of learning across varied performance environments (Ackerman, 1988; Shute, 1991). In this paradigm, individuals store old information while simultaneously processing new stimulus information.

The WM tasks in our battery required examinees to relate what was described in three statements to the order of four key words subsequently presented. For instance, one of the three statements might use category names (*animals, furniture*), whereas the other statements would use exemplars within those categories (*cow, chair*). Content complexity was incorporated through the use of faceting. An example of faceting in the WM–quantitative domain is the use of minus signs, which required participants to reverse mentally a given ordering of numbers. Multiple minus signs were also used, and these required several reversals.

Fact Learning

This domain was designed to assess declarative learning skills—that is, the ability to acquire new fact knowledge and commit that knowledge to memory. One test requires word pairs to be learned and then recalled after 10 pairs have been given. After they have been learned, one pair element is presented to prompt recall of the second element. In the quantitative domain, participants learned digit pairs. Another test required participants to learn 36 words. Participants then determined whether individually presented words were from the list. Complexity was incorporated through the use of limited-time learning. For example, only 4 s of learning time per pair might be permitted before participants were required to respond.

Skill Learning

This set of measures required procedural learning, the ability to acquire and use new rules and skills. For the verbal protocol, participants were shown a sentence consisting of three words, a noun, a verb, and an adverb. The participants were required to determine whether a displayed sentence conformed to one of two sentence patterns (that is, noun–verb–adverb or adverb–noun–verb). Inference was required in that only two of the three words were presented.

Induction

The IN measures involved the use of procedural knowledge and were an attempt to assess the depth and breadth of examinees' rule-centered knowledge base—that is, the ordering or unifying patterning of their learning. For example, in the verbal protocol, participants were shown three lists containing three nouns each. The participant then determined which of the three lists did not belong. For example, two lists might contain names of cities, and the third ("orphan") grouping might contain state names. In the spatial protocol, participants were shown a series of shapes and chose the next shape occurring in the series. Faceting was incorporated in that participants had to learn various themes to complete each trial successfully. One theme concerned additions being made to the initial figure, another involved rotation of figures.

The three stimulus domains that were crossed with the process factors are as follows: The *spatial* domain involves the rotation of objects, physical matching, and synthesizing of pictorial content. The *verbal* domain uses verbal operations such as linguistic transformations, category judgments, and part-of-speech classifications performed on words, sentences, and phrases. The *quantitative* domain uses quantitative operations such as arithmetic, sign reversals, and odd–even and high–low judgments on digits and numbers.

It is true that traditional, paper-based measures have somewhat of a limited scope because of their clearly restrictive ("low-tech") nature. According to Guilford's (1967, 1985) structure-of-intellect terms, an effective measure of intelligence might contain a variety of contents, require numerous distinct operations, and incorporate diverse end products. Furthermore, Humphreys (1981, 1984) has described the possibility of incorporating additional dimensions or facets in the measurement of ability, such as timing requirements and sensory modality. Accordingly, a primary advantage of the measures described here is that many different types of test stimuli and content are amenable to presentation. Finally, response production (i.e., through the use of partially open-ended questions requiring response generation) is required on several of the measures, which has only recently begun to be a realized possibility with traditional measures.

Data obtained from the intelligent tutoring systems (see Shute & Psotka, 1996; also Anderson, Boyle, & Reiser, 1985) both provided realistic performance environments and made available comprehensive outcome measures. Both tutors incorporated real-time information processing and mimicked the rapid decision making and problem solving required in many job situations, where performance environments are cognitively demanding. A wide range of job-relevant skills and knowledge was required on the tutors in content-integrated scenarios. Indeed, the tutors assessed many of the skills necessary in today's complex job situations, incorporating such knowledge facets as reasoning, declarative and procedural learning (Anderson, 1983), and skill acquisition. A brief description of the tutors follows, and more detail is given in a later section.

The electricity tutor teaches Ohm's and Kirchhoff's laws regarding current and voltage using electrical circuits.[2] Students are allowed to work on circuits to learn critical principles of electricity. The flight engineering tutor instructs participants in the use of tables, scales, and graphs, with the goal of making the calculations necessary for completing a take-off and landing data (TOLD) sheet.

Finally, the ASVAB, which is used to measure general ability, consists of the following subtests: General Science (GS), Word Knowledge (WK), Paragraph Comprehension (PC), Arithmetic Reasoning (AR), Math Knowledge (MK), Numerical Operations (NO), Coding Speed (CS), Auto and Shop Knowledge (AS), Mechanical Comprehension (MC), and Electronics Knowledge (EM). Armed Forces Qualification Test (AFQT; see Dunnette, 1976) scores, also used, were based on a linear combination of AR, MK, PC, and WK.

Study 1

Method

Participants. We obtained data for 344 participants who completed the ASVAB, the battery of computerized measures of cognitive information processing, and the computerized tutor designed to teach Ohm's law and other principles of electricity. The participants were employees from a local temporary service agency and were paid for their participation in the study. Fifteen percent of the participants were female, 39% were Caucasian, and 53% were members of minority groups.

Instructions were tailored to the specific tests. Sample items were given in each test with the instructions. Numbers of test items ranged from 20 in the induction tests to 192 in the SL–quantitative tests. The participants were given feedback as to the correctness of completed trials in each test, following individual responses or sets of responses. Finally, we established response goals for motivation and instructional purposes. For the processing speed tests, examinees were told that speed and accuracy were equally important for their scores. For the remaining tests, the participants were told that their scores would be dependent on the number they got correct but that they would have only a certain number of seconds to respond.

We used the cognitive process measures and the AFQT (a composite based on a subset of tests from the ASVAB) to predict learning proficiency and performance on the tutors. Because the tests of cognitive processing were timed examinations and were administered by computer, two summary scores were available for analysis: overall accuracy on test items (ACC) and median response time (RT) on items or trials averaged within the test. Response time latencies are believed to reflect, in part, the time required to retrieve critical information from long-term storage. Thus, the computerized cognitive-process measures were scored for both accuracy and latency at the item or trial level. Percentage of correct and median latencies, respectively, were used as test-level scores for analyses involving these measures.

We controlled speed–accuracy tradeoffs as much as possible. However, evidence of such was found, and, indeed, they are unlikely to be avoidable in the speeded-test domain. Some degree of control was afforded by our use of fixed item–time limits and by the feedback we gave on time to respond, as well as by our use of correct–incorrect error messages. Also, where appropriate, accuracy and latency feedback were given at the block level. Several tests used "timeouts" (limited times for problem solving or answer entry). Finally, motivation and debriefing screens were given when appropriate.

Participants were paid by the hour for their participation, which consisted of a number of hours of both testing and learning, the majority of which was spent using the tutor. The average time spent using the tutor was 11.5 hr (SD = 3.85). The influence of individual differences is clearly obvious as one examines the range of values for time to mastery—that is, 5.2–25.6 hr. Individuals were not told that their temporary employment as study participants would be a function of the time they spent on the tutor and would be hastened by their finishing the tutor. Thus, examinees were not differentially motivated to spend unnecessary time on the tutor. The cognitive-process measures required approximately 3 hr to complete; this included two 5-min breaks. Participants also spent approximately 3 hr on the ASVAB. Finally, we spread testing sessions out over a number of days to reduce measurement error from fatigue effects.

The participants were tested in groups of approximately 20 persons. The cognitive-process tests were administered on microcomputers with standard keyboards and monitors. The tutors were administered on Xerox 1186 computers with standard keyboards and high-resolution, monochromatic displays on 19 in. (48.26 cm) monitors. The participants viewed items on the display and responded by selecting a key on a standard computer keyboard.

The electricity curriculum consisted of 15 principles. Examinees were allowed to move to the next principle only after they solved three problems correctly on a given principle. Also, response histories were taken into account in determining types of problems to be presented; thus, an adaptive testing element was incorporated. Given the lengthy amount of training participants were given on the tutor, it can be said to be quite effective at replicating and emulating the complex learning and performance environments present in many job settings.

Participants were paid only after completing the entire study, which required mastery of the tutor knowledge and skills. Thus, we believed participants were motivated to finish the tutors in a timely and conscientious fashion. Given the importance of high levels of motivation in the maximum performance testing domain, such attention to detail clearly pays off in terms of low levels of measurement error.

Analyses. In addition to correlational analyses and covariance structure modeling, we performed regression analyses to examine the validity of the cognitive-process measures, as well as the unique contributions of the available measures beyond that provided by traditionally defined and assessed g. The dependent variables were time spent achieving mastery on the tutor and accuracy on the test of overall knowledge acquired from the tutor. The first criterion assessed how long participants took to acquire and master a thorough understanding of the principles of electricity. The second criterion measured the number of correct solutions and therefore assessed the participants' actual percentage correct on the final test rather than their discrete mastery alone.

To keep the number of variables in regression equations to a minimum to avoid unnecessary capitalization on chance, we used AFQT scores to represent the general factor of intelligence. Given that the AFQT has long been used for selection and placement decisions in the military and has received considerable psychometric, measurement, and validation attention and that AFQT scores correlated .94 with a linear composite of the ASVAB scores in the current data, this decision seemed to be well justified (see also Wigdor & Green, 1991). Furthermore, it is important to note that sample means and standard deviations for all of the ASVAB subtests were quite close to the means and standard deviations of the ASVAB normative sample—that is, means of approximately 50 and standard deviations of approximately 10.[3] Thus, the sample had at least 90% of the variability of the normative sample. This finding indicates minimal restriction of range on relevant score distributions and also points to the very adequate and

[2] Ohm (Ω) refers to Georg Simon Ohm (1787–1854), a German physicist who showed that an ohm (a meter–kilogram–second unit of resistance) is equal to the electric resistance in a conductor in which one volt of potential difference produces a current of one ampere. Ohm's law states that voltage = current × resistance.

[3] AFQT has been calibrated on a national probability sample of 18- to 23-year-olds (i.e., those who are eligible for service).

even excellent motivation levels present for nonoperational testing. Finally, we carried out analyses using individual ASVAB scores in place of AFQT or ASVAB composite scores. Highly consistent results were obtained.

Because of the large number of tests in the cognitive-process battery, the number of variables in regression equations was further minimized. First, the tests were grouped into composites on the basis of the six cognitive-process factors measured in the battery: PS, WM, FL, SL, IN, and GK. We formed composites by adding together, using unit weighting, scores from the quantitative, verbal, and spatial domains within each of the factors. Thus, six composites were formed from the mean accuracy scores, which were essentially percentage-correct scores. These six composites were entered simultaneously into regression equations.

Because participants were to answer individual test items quickly and also had item time limits, a second set of composites used RT scores. We again formed composites on the basis of the six factors, in the same manner we used to form the ACC composites.

Results and Discussion

Table 1 contains reliabilities and correlations among cognitive-task latency scores and accuracy scores (within score type only). Because both studies used the same cognitive tests, it was possible to compute correlations and reliabilities on the basis of the combined data ($N = 746$) to estimate values in the most stable fashion. We believed this decision was reasonable, because samples were similar in terms of make up, and performance was similar across samples. It is important to note that correlations are almost entirely nonzero and positive. The average intercorrelation for RT scores (above the diagonal) was .20, and the average correlation for ACC scores was .29. Alphas were mostly quite high, with the exception of those for the GK–quantitative domain ACC scores. It is relevant to note that the consistency in the correlations provides initial evidence for a general latent factor driving the multiplicity of responses.

Table 2 contains correlations between tests of the two score types. It is important to note that these correlations were quite a bit lower overall than the within-score correlations were. They seem to reveal some, albeit a limited, degree of commonality among accuracy scores and speed of responding.

Table 3 contains correlations between the cognitive task scores and the two electricity criterion scores, as well as the overall AFQT scores. It is important to note the direction of the accuracy score–total time correlations; the direction indicates that longer times were associated with lower accuracy scores and vice versa. In general, accuracy scores appear to perform better in a correlational sense than do the latencies (as Humphreys, 1989, anticipated some time ago). These results are in keeping with Kyllonen's (1994) work, which reported correlations between ASVAB test scores and various cognitive process accuracy scores in the range of .56–.96.

Table 4 contains zero-order correlations between the general ability measure and the two tutor-performance scores. Also, unit-weighted composites of all cognitive-task scores (separately for latencies and accuracies) were correlated with the various criteria and other relevant scores. The correlation between the two tutor criterion scores is −.65. It is important to note the considerable degrees of commonality between cognitive measure accuracy scores and AFQT and tutor criterion scores.

Next, we performed a hierarchical confirmatory factor analysis (HCFA) to examine the theoretical common factor structure of the various test batteries. We hypothesized a model positing three higher order (second-order) factors, one for each of the three types of measures assessed here (traditional *g*, cognitive-process accuracy *g*, and cognitive-process latency *g*). Then we specified three first-order factors for the ASVAB test scores. We used the entire ASVAB battery (10 subtest scores), because this allowed estimation of latent structure, whereas using a single AFQT score would not. Ree and Carretta (1994) factor analyzed the ASVAB using confirmatory factor analysis and found support for a hierarchical structure resembling Vernon's (1969) theoretical specifications. The same grouping of tests that Ree and colleagues employed was used here (Ree & Carretta, 1994; Stauffer, Ree, & Carretta, 1996). That is, we hypothesized a verbal–mathematical factor (V-M) consisting of the AR, WK, PC, and MK subtests. We also posited a speed factor (SPD; derived from NO and CS) and a technical knowledge factor (TK; derived from GS, AS, MC, and EM). We posited that these three first-order factors would load on the first of the 3 second-order factors (g_A) representing traditionally constituted general ability.

Next, we grouped the cognitive process measures by test domain (PS, WM, FL, SL, IN, GK) and score type (accuracy and latency), yielding 12 first-order factors. (We also carried out grouping by content domain, which yielded a highly comparable fit.) Six factors were allowed to load on each of the two higher order factors, one general factor for the accuracy score (g_{ACC}) and one general factor for the latency scores (g_{RT}). Also, as Humphreys (1989) noted, variables that negatively correlate cannot by definition load on a general factor. Accordingly, latency scores were reflected for the factor analytic work. Finally, the three general (second-order) factors (g_A, g_{ACC}, g_{RT}) were allowed to load on a single, general (third-order) factor representing general intelligence (*g*).

Figure 1 shows the hierarchical factor model along with path loadings for the first-order factors loading on the 3 second-order factors and for these factors loading on the third-order *g* factor. Loadings of observed variables on the first-order factors (not shown) were as follows: ASVAB tests (.723–.916), ACC scores (.143–.827, with PSV—the compilation of processing speed and the verbal domain—showing an outlying value of .143 and the next lowest value being .309), and RT scores (.226–.926). All loadings were significant at $p < .001$. Root mean square error of approximation (RMSEA; Browne & Cudeck, 1993) was .069 for this model (with a confidence interval of .067–.072), indicating very good to excellent fit. It is important to note the generally high levels of loadings of first-order factors on the second-order *g* factors (g_A, g_{RT}, and g_{ACC}). Also relevant is the fact that the path estimates between the second-order factors and the third-order general factor are indicative of considerable commonality among the three latent general factors. The somewhat disappointing path for g_{RT} may be due to speed–accuracy trade offs and other psychometric concerns that would affect the latency metric unfavorably. These results point to additional confirmation of the notion of a theoretical general factor that is even more broadly constituted than we previously believed.

We used Marsh and Hocevar's (1985) target coefficient to compare the hierarchical (third-order general factor) model analyzed here with other possible models. We analyzed a full first-order model wherein the 15 first-order factors were allowed to load on a single second-order factor. The RMSEA for this model was .073 (.071–.076). The favored hierarchical model, being much

Table 1
Reliabilities and Intercorrelations of Cognitive-Task Latency and Accuracy Scores

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. PSV	.73/.93	.18	.24	.24	.26	.09	.30	.04	.08	.21	.16	.03	.06	.16	.03	−.06	−.01	.07	.07
2. PSQ	.15	.89/.95	.30	.23	.22	.17	.21	.16	.23	.31	.27	.16	.13	.21	.10	.05	.09	.14	.23
3. PSS	.08	.26	.79/.94	.24	.19	.15	.21	.12	.19	.26	.26	.10	.06	.25	.10	.14	.13	.16	.22
4. WMV	.15	.36	.44	.96/.87	.29	.21	.27	.15	.14	.24	.19	.07	.05	.23	.21	.15	.13	.21	.17
5. WMQ	.14	.37	.42	.56	.92/.91	.25	.28	.17	.22	.22	.11	.12	.14	.20	.11	.07	.08	.18	.14
6. WMS	.15	.33	.47	.58	.57	.89/.88	.17	.14	.38	.27	.20	.14	.11	.18	.23	.20	.24	.26	.21
7. FLV	.21	.29	.24	.35	.36	.40	.92/.88	.17	.19	.23	.20	.07	.03	.21	.11	.07	.06	.14	.04
8. FLQ	.12	.21	.33	.34	.32	.40	.33	.80/.93	.31	.19	.24	.12	.08	.06	.12	.14	.14	.15	.26
9. FLS	.07	.25	.43	.52	.48	.59	.32	.36	.84/.71	.32	.21	.17	.14	.20	.22	.19	.23	.23	.29
10. SLV	.09	.29	.44	.41	.49	.46	.40	.37	.35	.95/.70	.35	.24	.16	.27	.11	.13	.13	.16	.26
11. SLQ	.12	.39	.42	.54	.56	.52	.39	.39	.48	.52	.96/.70	.15	.12	.12	.09	.14	.09	.08	.22
12. INV	.05	.28	.31	.39	.32	.32	.20	.25	.31	.30	.37	.73/.82	.42	.37	.15	.18	.16	.20	.29
13. INQ	.12	.38	.40	.44	.48	.50	.28	.29	.42	.38	.44	.35	.77/.79	.25	.10	.08	.06	.06	.19
14. INS	.04	.29	.41	.50	.46	.50	.26	.30	.46	.35	.51	.41	.42	.71/.78	.19	.13	.16	.20	.17
15. GKQ1	.00	.02	.06	.10	.06	.07	.02	.02	.15	.03	.09	.11	.13	.08	*/.80	.75	.69	.68	.15
16. GKQ2	−.03	.09	.13	.11	.07	.14	.01	.13	.15	.07	.12	.12	.11	.17	.16	*/.86	.72	.70	.17
17. GKQ3	−.02	.25	.25	.33	.28	.34	.15	.19	.35	.20	.32	.35	.34	.36	.27	.27	.46/.83	.74	.16
18. GKQ4	−.03	.21	.14	.24	.16	.18	.07	.07	.18	.14	.16	.25	.23	.23	.23	.18	.36	.38/.92	.17
19. GKS	.07	.30	.43	.50	.48	.51	.33	.32	.50	.38	.49	.40	.42	.49	.06	.10	.42	.21	.84/.88

Note. Latency correlations are above the diagonal; accuracy correlations are below. Accuracy reliabilities are the first in the pairs on the diagonal; latency reliabilities are the second. $N = 648$. $rs >$ .10 are significant at $p < .01$. Asterisks indicate that the reliability is unavailable. Latency and accuracy measures (and the abbreviations for these measures) are composites of the following factors and domains: PS = processing speed; WM = working memory; FL = fact learning; SL = skill learning; IN = induction; GK = general knowledge; V = verbal; Q = quantitative; S = spatial domain.

Table 2
Correlations Between Cognitive-Task Accuracy and Latency Scores

	Latency score																		
Accuracy score	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. PSV	.15	.09	.05	−.03	.00	.00	.03	.09	.06	.13	.08	.12	.05	.09	−.03	−.01	.03	−.02	.09
2. PSQ	−.18	−.14	−.02	−.13	−.27	−.12	−.09	.05	−.04	−.07	.07	.04	−.02	.00	−.03	.04	.02	−.06	.03
3. PSS	−.24	.01	.08	−.16	−.22	−.02	−.21	−.01	.08	−.03	−.00	.10	.11	−.05	.05	.09	.08	−.01	.14
4. WMV	−.38	−.09	−.07	−.36	−.40	−.08	−.30	−.01	.03	−.09	.03	.06	.05	−.13	.02	.09	.06	−.10	.02
5. WMQ	−.35	−.00	−.02	−.23	−.43	.01	−.29	.03	.03	.03	.14	.11	.10	−.03	.02	.07	.07	−.04	.08
6. WMS	−.37	−.09	−.08	−.29	−.34	.08	−.32	.03	.02	−.08	.00	.10	.04	−.11	.02	.09	.10	−.02	.05
7. FLV	−.25	.05	.00	−.13	−.16	.03	−.06	.09	.08	.07	.12	.07	.06	.02	.01	.06	.01	−.07	.07
8. FLQ	−.33	−.01	−.05	−.15	−.18	.02	−.19	.26	.07	−.00	.05	.12	.02	−.04	.03	.10	.06	−.04	.11
9. FLS	−.36	−.09	−.13	−.23	−.32	−.05	−.29	.04	.03	−.15	−.00	.08	.05	−.19	.03	.11	.08	−.02	.02
10. SLV	−.31	.08	.03	−.12	−.26	.07	−.14	.04	.08	.04	.13	.13	.14	.01	.04	.12	.10	−.04	.06
11. SLQ	−.31	.02	.03	−.19	−.26	−.02	−.20	.08	.04	.05	.37	.14	.11	−.05	.03	.10	.06	−.04	.13
12. INV	−.19	−.12	−.05	−.20	−.24	−.12	−.23	.03	−.05	−.09	.01	.06	.11	−.06	−.02	.06	.05	−.04	.06
13. INQ	−.25	−.05	−.02	−.15	−.28	−.06	−.17	.05	−.00	−.01	.06	.19	.22	.05	.03	.09	.05	−.04	.07
14. INS	−.29	.06	−.04	−.20	−.27	−.06	−.19	.02	−.01	−.08	.11	.22	.17	.01	.07	.09	.08	−.03	.08
15. GKQ1	−.09	−.09	−.08	−.05	−.07	−.02	−.05	.03	.00	−.05	−.04	−.02	−.01	−.05	−.01	−.00	.03	−.04	−.02
16. GKQ2	−.09	−.08	−.01	−.06	.08	−.11	−.04	−.04	−.07	−.11	−.05	−.02	−.05	−.04	.02	.06	.03	.01	−.04
17. GKQ3	−.23	−.22	−.09	−.19	−.26	−.17	−.29	−.05	−.12	−.19	−.07	−.01	−.05	−.09	.01	.13	.07	−.05	−.08
18. GKQ4	−.12	−.17	−.05	−.13	−.12	−.06	−.15	−.04	−.07	−.15	−.06	−.02	−.01	−.08	.05	.09	.05	−.03	−.05
19. GKS	−.31	−.05	−.07	−.21	−.35	−.07	−.26	.03	−.04	−.10	.01	.12	.04	−.12	.01	.09	.07	−.07	.21

Note. $N = 648$. $rs > .10$ are significant at $p < .01$. Latency and accuracy measures (and the abbreviations for these measures) are composites of the following factors and domains: PS = processing speed; WM = working memory; FL = fact learning; SL = skill learning; IN = induction; GK = general knowledge; V = verbal; Q = quantitative; S = spatial domain.

Table 3
Correlations Between Cognitive-Process Scores and Electricity Concepts Tutor and AFQT Scores

	Accuracy scores			Latency scores		
Measure	Total time on tutor	Posttutor accuracy	AFQT	Total time on tutor	Posttutor accuracy	AFQT
PSV	−.09	.09	.12	.36	−.33	−.38
PSQ	−.40	.35	.51	.12	−.06	−.06
PSS	−.47	.47	.48	.08	.00	.01
WMV	−.56	.49	.65	.36	−.24	−.33
WMQ	−.45	.46	.59	.42	−.30	−.41
WMS	−.54	.51	.59	.19	−.12	−.16
FLV	−.26	.35	.37	.34	−.34	−.36
FLQ	−.23	.30	.33	.11	.04	−.03
FLS	−.43	.55	.59	.17	−.05	−.13
SLV	−.43	.41	.49	.13	−.03	−.12
SLQ	−.51	.54	.64	−.01	.09	.06
INV	−.38	.43	.57	−.03	.08	.04
INQ	−.41	.47	.56	−.06	.14	.10
INS	−.54	.52	.61	.09	−.04	−.09
GKQ1	.04	.04	.06	.08	−.05	−.13
GKQ2	−.14	.08	.18	.04	.05	−.03
GKQ3	−.43	.49	.53	.01	.02	−.04
GKQ4	−.13	.18	.30	.13	−.06	−.18
GKS	−.53	.53	.64	.04	.06	.05

Note. $N = 292$. $rs > .14$ are significant at $p < .01$. Latency and accuracy measures (and the abbreviations for these measures) are composites of the following factors and domains: PS = processing speed; WM = working memory; FL = fact learning; SL = skill learning; IN = induction; GK = general knowledge; V = verbal; Q = quantitative; S = spatial domain. AFQT = Armed Forces Qualification Test.

simpler, had 87 fewer parameters than the full second-order factor model did. In terms of parsimony, the hierarchical model has much to recommend it. The target coefficient, whose upper limit is 1.0, was .81, indicating that a substantial proportion (81%) of the relations among the first-order factors was accounted by the higher order factors (Marsh & Hocevar, 1985).

Another main rival model was one with 3 second-order correlated factors representing the same general factors analyzed in the favored model; an RMSEA of .068 (.066–.070) was derived for this rival model. The favored model had a slightly worse fit but was much more appealing theoretically. The higher order factor explains the correlations among the lower order factors in a more parsimonious and appealing way (Marsh & Hocevar, 1985). Indeed, Browne and Cudeck (1993) stated, "Model selection has to be a subjective process involving the use of judgment" (p. 57). Fit indices tell us little about a model's plausibility (Browne & Cudeck, 1993). It is also important to note that this "model" is inadequate in that it does not explain the correlations among the 3 second-order factors.

Correlations among the 15 first-order factors for the favored third-order hierarchical model (these are not shown) ranged from .114 to .996, with the majority of correlations falling much closer to the higher value. Furthermore, as Stauffer, Ree, and Carretta (1996) have also reported, the largest correlations between first-order factors defining cognitive-process scores and traditionally defined *g* measures (g_A measures) occurred for V-M (the verbal–mathematical factor), which the authors noted is an excellent "avatar of '*g*' " (p. 199). Alternatively, the largest correlations involving cognitive-process scores and other factors were found for PS_{ACC} (and also for IN_{ACC}). Indeed, the correlation between V-M and PS_{ACC} was .913 ($r = .909$ for IN_{ACC}). We carried out multigroup analyses comparing factor structure for men and women and for Whites and minority group members. RMSEAs were .05 (.050–.053) for both analyses, indicating excellent levels of fit and high degrees of comparability for the general latent factor structure across gender and race.

Lastly, we carried out analyses to estimate the strength of the theoretical relationship between the newly constituted general (third-order) factor from the HCFA and a factor defining criterion scores. Comparisons can be made between the path estimate for traditional g_A alone (the ASVAB–AFQT factor) and the general factor made up of all three types of scores from the HCFA reported previously (*g*). For the tutor accuracy criterion, the path from traditionally constituted g_A was .823, and that from the newly

Table 4
Correlations Among Tutor Scores, AFQT Scores, and Composites of Cognitive-Process Scores

Measure	1	2	3	4	5
1. AFQT	—				
2. Tutor accuracy	.73	—			
3. Tutor time	−.69	−.65	—		
4. Process accuracy	.78	.85	−.68	—	
5. Process latency	−.21	$-.08_a$	.22	$.05_a$	—

Note. $N = 322$. All correlations $p < .01$ except as indicated by subscripts. AFQT = Armed Forces Qualification Test.

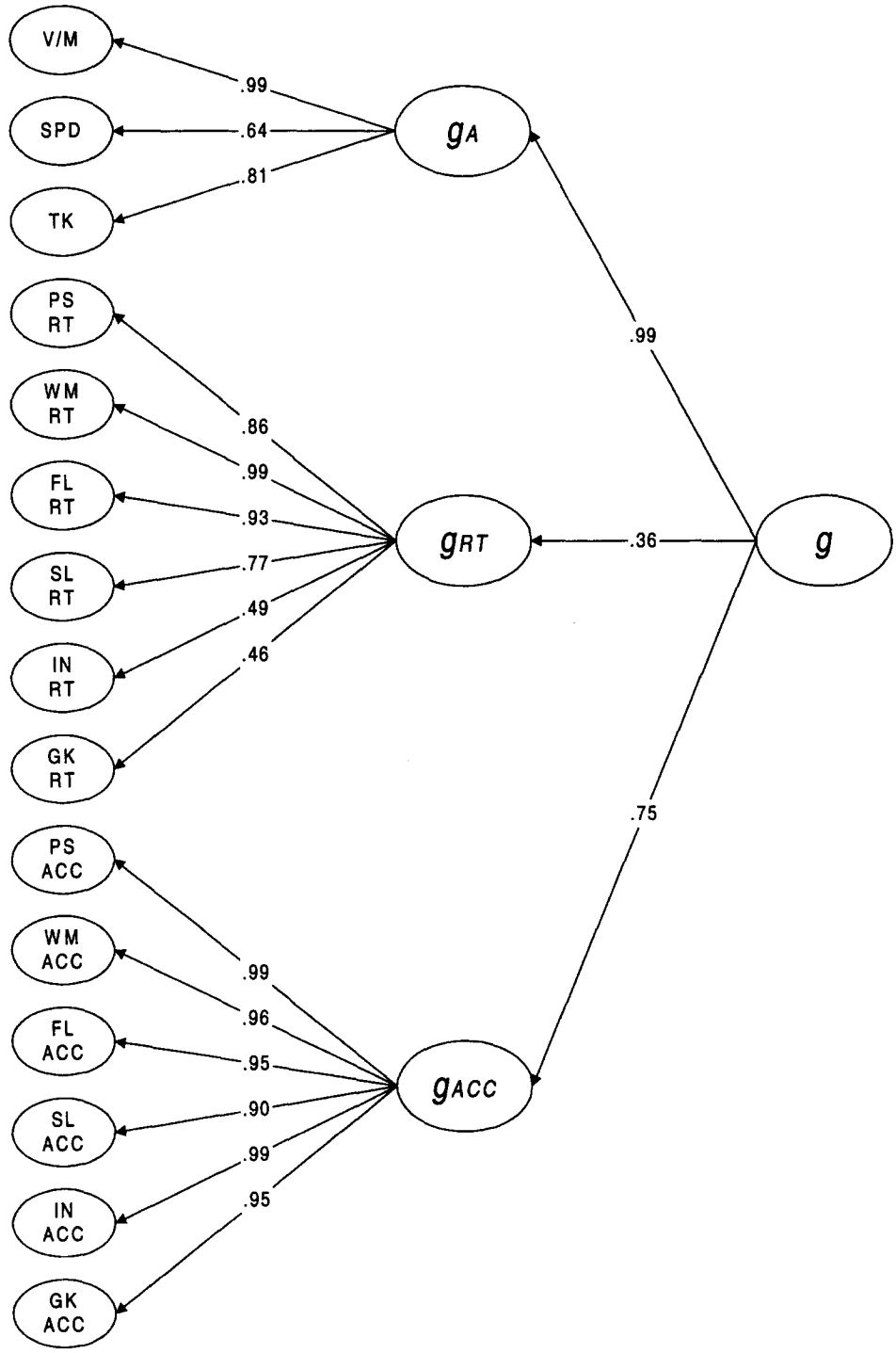

Figure 1. Third-order factor model with maximum likelihood estimates of path loadings. Working memory (WM), general knowledge (GK), processing speed (PS), fact learning (FL), skill learning (SL), and induction (IN) represent the cognitive-process measures. RT represents reaction time, and ACC refers to accuracy. V/M (verbal/mathematical), SPD (speed), and TK (technical knowledge) are factors from the Armed Forces Qualification Test (AFQT), a composite based on a subset of tests from the Armed Services Vocational Aptitude Battery (see Dunnette, 1976). g_A = the second-order factor reflecting AFQT scores; g_{ACC} = the second-order factor reflecting ACC scores; g_{RT} = the second-order factor reflecting latencies; g = the third-order, general factor.

Table 5
Validity and Incremental Validity of AFQT, Accuracy (ACC), and Latency (RT) Composites for Accuracy on Electricity Concepts Tutor Posttest

Correlation	AFQT	ACC	RT	ACC added to RT	RT added to ACC	Added to AFQT: ACC	Added to AFQT: RT	Added to AFQT: RT and ACC
R	.73*	.74*	.38*	.74*	.74*	.76*	.74*	.77*
R^2	.53*	.55*	.15*	.55*	.55*	.58*	.54*	.59*
Adjusted R^2		.54*	.13*	.53*	.53*	.57*	.53*	.57*
Δ Adjusted R^2				.38*	.00	.05*	.00	.04*

Note. N = 322. AFQT = Armed Forces Qualification Test.
*$p < .01$.

constituted g was .913. These paths were −.766 and −.772, respectively, for the time-to-mastery criterion. Both analyses represent improvements for the addition of the factor space representing cognitive-process scores.

Next, we carried out regression analyses to determine the comparative utility of the various test and score batteries, as well as the incremental value of the cognitive-process test scores. Table 5 presents results of regression analyses using the RT and ACC composites and AFQT scores to predict performance on the posttutor accuracy measure. Table 5 reports validities for these scores for posttutor accuracy, as well as the incremental validities obtained when a cognitive-process composite type was added to the other cognitive composite type or to the AFQT in the regression equation. All regression tables include squared multiple correlations adjusted for the differing numbers of predictors representing the accuracy, latency, and traditional g scores. The shrinkage formula producing the best estimators of squared multiple correlations for the population determined by the computed sample estimates was used to provide as complete and accurate a comparison with sample results as possible (Cattin, 1980).

The ACC composite resulted in the largest squared multiple correlation for predicting posttutor accuracy (54% of the variance), followed by the AFQT (53%), and then RT scores (13%). Cognitive test ACC scores added unique variance to both RT and AFQT scores (38% and 5%, respectively). Moreover, ACC scores made a unique contribution to the predictability of criterion variance beyond both of these latter scores combined (5%; not shown in Table 5). Clearly, the cognitive-process error rate (ACC) scores resulted in highly useful measures, both on their own and as an addition to the traditionally defined and measured assessment of general ability. It is noteworthy, however, that the traditional g composite measure also showed incremental validity over the cognitive-process measures, adding unique variance beyond the ACC and RT composites combined (4%; not shown in Table 5). Thus, both classes of measures appear to be useful. On the other hand, RT on the computerized information-processing measures is not particularly effective on its own in predicting variance over that accounted for by ACC scores or scores from traditional measures. However, the linear combination of latency and ACC scores clearly provides additional variance beyond traditional g (4% total).

We carried out similar analyses for the time-to-mastery criterion. Table 6 presents the results of regression analyses using total time to complete the tutor program as the dependent variable. As we observed for the first criterion, ACC had the largest overall squared multiple correlation (adjusted .48; R = .70). That is, accuracy alone explained 48% of the variance in the total time to achieve mastery on the tutor program. AFQT explained 46% of the variance in time spent on the tutor, and RT alone accounted for 24%. It is important to note that RT and ACC added 8% to the AFQT scores. Given that RT is essentially a time-based measure, one might expect that it would account for unique variance in time spent on the tutor. Indeed, RT added 4% unique variance to ACC and AFQT combined (not shown in Table 6). On the other hand, ACC again added the most unique variance beyond the others combined (6%; not shown).

As in the analyses with the posttutor accuracy criterion, AFQT scores also contributed unique variance beyond ACC and RT in predicting the time-to-mastery criterion, though the variance added was small (1%). Thus, in both analyses, the cognitive-process tests

Table 6
Validity and Incremental Validity of Accuracy (ACC) and Latency (RT) Composites for Total Time on Electricity Concepts Tutor

Correlation	AFQT	ACC	RT	ACC added to RT	RT added to ACC	Added to AFQT: ACC	Added to AFQT: RT	Added to AFQT: RT and ACC
R	.68*	.70*	.50*	.74*	.74*	.72*	.72*	.75*
R^2	.46*	.49*	.25*	.55*	.55*	.52*	.51*	.57*
Adjusted R^2		.48*	.24*	.53*	.53*	.51*	.50*	.55*
Δ Adjusted R^2				.28*	.04	.05*	.04*	.08*

Note. N = 322. AFQT = Armed Forces Qualification Test.
*$p < .01$.

added more unique variance to AFQT than AFQT added to them in predicting the criteria of interest, and accuracy on the cognitive process measures was the best overall predictor, though by only a small margin. How might one interpret these results? Simply, the cognitive-process measures alone contained more useful variance in predicting the criterion than did the AFQT scores.

A means of keeping the number of variables in the regression equations at a reasonable level was to select only the very best of the information-processing tests for analysis. We performed this analysis to determine whether comparable validity and construct space coverage was achievable using fewer tests. Tests were selected on the basis of their measurement and psychometric properties, including reliabilities, correlations with measures of ability, and item difficulty distributions. The majority of the CAM tests were chosen for the ACC-based scores (only General Knowledge Quantitative 1 and 2 and Processing Speed Verbal were eliminated from analyses). However, only six latency-based scores were chosen (Processing Speed Verbal, Working Memory Verbal, Working Memory Spatial, Working Memory Quantitative, Fact Learning Verbal, and Induction Quantitative). It is important to note that all of the working memory tests were "tagged" as highly useful from a psychometric standpoint.

Results were similar to those using composites of all scores. The squared multiple correlations for these analyses were found to be of a greater overall magnitude for the ACC and RT scores for the selected tests (most gains were around 2% beyond the R^2 gains presented previously). The notable exception was the RT scores, which performed much better under the selection method than using composites (R^2 = .25 vs. .15). In addition, although the R^2s or ACC, RT, and AFQT were somewhat larger when the best tests were used, we found the trends among them to be highly comparable with those reported for the overall composites.

Regression analyses using scores from the selected tests were also carried out for the time-to-skill mastery criterion. Improvements were realized and were similar to those for the final accuracy test. The tests showed an increase in squared multiple correlation for time to mastery, usually of around 4%, when the psychometric selection method was used instead of overall composites. Similar trends were observed for the time criterion. That is, ACC had the largest squared multiple correlation, followed by AFQT and then RT.

Study 2

Method

In Study 2, 402 participants completed the ASVAB battery, the computerized cognitive measures, and a tutor designed to teach flight engineering knowledge and skills. As in the first study, participants were temporary service employees. Approximately 25% of the participants were female, 35% were Caucasian, and 61% were members of minority groups.

The flight engineering tutor was originally developed by researchers at the University of Pittsburgh (Lesgold, Busszo, McGinnis, & Eastman, 1989) and was subsequently modified at the Air Force's Armstrong Laboratory (Shute, 1993b). The tutor was designed to simulate the learning of skills and knowledge associated with a flight engineer's job. Relevant components included collecting and analyzing information about a pending flight and determining whether various factors (e.g., type of plane, current weather) would allow a safe flight. The tutor had two main parts: reading graphs and completing a TOLD sheet. Problem sets were included for both parts of the tutor. All problem sets involved the learning of procedural and declarative knowledge and contained problems pertaining to scale conversions, linear relationships, and interpreting polar and Cartesian coordinate charts.[4]

An example of procedural knowledge requirements for the tutor is computing relative wind direction. Other requirements called for participants to figure out the maximum allowable crosswind given the gross weight of an aircraft and runway conditions. In a step-by-step manner, the tutor demonstrated how to determine allowable crosswind before problems were presented.

The vast majority of the material in the tutor curriculum constituted new knowledge and skills for the examinees. Graph reading, for instance, consisted of 14 problem sets that required reading and understanding graphs, beginning with simpler graphs and progressing to complex ones. TOLD requirements consisted of 9 problem sets and involved learning and using graph reading skills to fill out the final TOLD worksheet. Participants were required to integrate procedures with the given information to determine whether conditions were acceptable for a safe takeoff or landing. For the work sheets, some information was given and other information had to be computed by the examinees. The procedure to determine headwind and crosswind was a detailed eight-step procedure. Computerized help was available to assist examinees. Finally, all learning was self-paced, which allowed the participants as much time as needed to master the material.

We again conducted regression analyses, using the overall amount of information learned from the tutor as the dependent variable. This criterion was a measure of the percentage correct on a comprehensive mastery test given following tutor completion.

Results and Discussion

Table 7 contains correlations between both types of cognitive-task scores and the posttutor accuracy measure, as well as AFQT scores. The zero-order correlation between the single tutor criterion score and the AFQT was .83. It is important to note the generally high level of commonality for ACC measures with both criterion and AFQT scores. Latencies again fared less favorably overall. Exceptions are the PS, WM, and FLV tests, as was the case in the first study.

Tables 8 summarizes the results of regression analyses, again using the ACC and RT composites. The same tests used for composites in Study 1 were used for these analyses. Here, the AFQT and the processing tests accounted for considerably more variance than they did in Study 1 (83% of posttutor mastery in Study 2, vs. 57% of posttutor mastery and 55% of time to mastery in Study 1).

Although there is a difference in overall magnitudes of the squared multiple correlations, the incremental-variance-accounted-for results indicate trends that are highly similar to those we observed for the composites in Study 1. In both cases, ACC had the highest squared multiple correlation; followed by AFQT and then the RT composite. Both ACC and RT scores showed incremental validity over each other and over AFQT, as in the analysis with the total time criterion in the first study. It is important to note that ACC contributed 9% unique variance beyond AFQT and that RT and ACC added 11% unique variance over AFQT. Results indicate that the variance added by RT over the percentage correct and AFQT combined was small but significant (2%; not shown in

[4] Cartesian coordinates are a system of coordinates for locating a point on a plane by the length of its radius vector and the angle this vector makes with a fixed line.

Table 7
Correlations Between Cognitive-Task Scores and Flight Engineering Tutor and AFQT Scores

	Accuracy		Latency	
Measure	Posttutor accuracy	AFQT	Posttutor accuracy	AFQT
PSV	.02	.00	−.42	−.38
PSQ	.41	.36	−.27	−.33
PSS	.52	.45	−.18	−.25
WMV	.60	.56	−.32	−.32
WMQ	.59	.53	−.44	−.51
WMS	.71	.60	−.10	−.23
FLV	.34	.23	−.36	−.35
FLQ	.45	.40	.02	−.01
FLS	.58	.53	.06	−.07
SLV	.47	.38	−.24	−.27
SLQ	.59	.51	−.03	−.10
INV	.53	.51	.24	.16
INQ	.59	.51	.15	.05
INS	.62	.57	−.19	−.25
GKQ1	.20	.24	.06	−.00
GKQ2	.15	.18	.19	.14
GKQ3	.55	.57	.12	.07
GKQ4	.35	.36	−.05	−.11
GKS	.68	.68	.15	.44

Note. $N = 402$. $rs > .14$ are significant at $p < .01$. The correlation between AFQT and posttutor accuracy is .83. AFQT = Armed Forces Qualification Test. Accuracy and latency measures (and the abbreviations for these measures) are composites of the following factors and domains: PS = processing speed; WM = working memory; FL = fact learning; SL = skill learning; IN = induction; GK = general knowledge; V = verbal; Q = quantitative; S = spatial domain.

Table 8). ACC scores again added the largest amount of unique variance over RT and AFQT combined (approximately 8%; not shown in Table 8), whereas AFQT added 6% unique variance over the others.

Finally, covariance structure modeling was again carried out to compare the size of the path estimate for the newly constituted general factor (g) and that for the traditionally composed factor (g_A). These path estimates were .98 and .91, respectively, predicting the theoretical estimate of the criterion of tutor accuracy. Thus, the construct defined by the cognitive process tests along with the traditional measures again represents an improvement in measurement of the latent factor space and predictability of the criterion factor space.

General Discussion

This research has succeeded in accomplishing several things. Results indicate that both accuracy- and reaction time-based scores from cognitive-process measures are highly useful for predicting learning and performance on complex tutors. Furthermore, these test scores contributed unique variance beyond that provided by a standardized measure of general ability. In two studies, regression analyses indicate that cognitive-process accuracy scores accounted for significantly more overall variance in criterion scores than that accounted for by traditionally defined and constituted general intelligence measures. Response latency scores alone added unique variance, but scores reflecting error rates from the cognitive-process measures were consistently the most useful predictors, whether they were used alone or in addition to scores from other tests. Finally, and also important, the cognitive-process measures, along with the traditional measures, appear to reflect a single higher order factor indicating general intelligence that also does an excellent job predicting criterion scores.

Christal (1991) carried out similar research, predicting the performance of military recruits on a computerized tutor designed to teach knowledge and skills on "logic gates" using a similar battery of experimental cognitive tasks and the ASVAB. He found that the cognitive-process battery added as much as 22% unique variance over ASVAB scores. The present study found up to 11% unique variance of experimental measures over traditionally constituted g. However, it is important to note that in the current research, less variance remained after the g measure had been entered into the regression equation than remained in Christal's study after test scores had been entered. The correlation between the general ability measure and a composite of the accuracy measures from the computer battery was much lower in Christal's study (e.g., .48 vs. .78 in the present research), thus allowing potentially more predictable variance to remain. Thus, it is noteworthy that incremental validities were obtained for the ACC scores in this research, in spite of the high initial validities for the AFQT.

We now move to a discussion of the broader implications of these results, both the regression analyses and the HCFA. The relationship between cognitive ability and job performance is well established and widely known (Hunter, 1986; Hunter & Hunter, 1984; Ree & Earles, 1991, 1992; Schmidt & Hunter, 1998; Wigdor & Green, 1991). Much work has been carried out that asks how much criterion variance is due to g versus how much is due to other non-g factors such as personality characteristics (e.g., Hunter & Hunter, 1984). However, we have attempted to reframe and

Table 8
Validity and Incremental Validity of Accuracy (ACC) and Latency (RT) Composites for Accuracy on Final Test of Flight Engineering Tutor Knowledge

						Added to AFQT		
Correlation	AFQT	ACC	RT	ACC added to RT	RT added to ACC	ACC	RT	RT and ACC
R	.85*	.87*	.56*	.88*	.88*	.91*	.87*	.91*
R^2	.72*	.76*	.32*	.77*	.77*	.82*	.75*	.83*
Adjusted R^2		.75*	.30*	.76*	.76*	.81*	.75*	.83*
Δ Adjusted R^2				.45*	.01	.09*	.03*	.11*

Note. $N = 402$. AFQT = Armed Forces Qualification Test.
* $p < .01$.

redirect this question by expanding the standard definition and operationalization of general ability. Humphreys and Lubinski (Humphreys, 1979; Lubinski & Dawis, 1992; Lubinski & Humphreys, 1990, 1992) have discussed the need for, and the possibility of, doing just that; their research indicates that the general factor is indeed quite a bit broader and richer than is operationalized in most current, standardized measures. Again, Humphreys' (1971) definition of intelligence as the repertoire of knowledge and skills in the cognitive domain available to a person at a particular point in time incorporates the idea that facility in acquiring new skills and performing on tasks requiring those skills is part of the general factor and should be contained in measures of the general factor. Indeed, the various cognitive-process measures investigated here can each be viewed as individual, narrow indicators or as reflections of the broadly constituted general factor (Detterman, 1982; Humphreys, 1971). It is possible that with the studied battery, we have achieved a different mix and thereby coverage of the actual latent factor space defining intelligence. However, such diverse reflections are rarely, if ever, present in standard predictor assessments. We believe this research has effectively begun the critically important process of more carefully and completely mapping the relevant construct space of general ability.

The generally very high levels of predictive efficiency, along with the theoretical path estimates for the combined batteries, are compelling. We do very well with the combined batteries in explaining variance on the realistic tutors. These findings are made all the more impressive on consideration of the fact that we started out with an established battery (the AFQT) that is of excellent psychometric quality. Given that the tutors were designed to mirror learning and performance environments in real-world jobs, we believe that these results bode well for generalizations to performance in actual workplace settings. On the tutors examined, novelty precluded ceiling effects, and careful instruction and self-pacing prevented serious floor effects. With the help of the computerized cognitive-process measures, we clearly exceeded the $r_{xy} = .50$ prediction barrier discussed by Lubinski and Dawis (1992). As a matter of fact, we have even handily eclipsed a .50-variance-accounted-for barrier. It is true that our criteria are lab-based and are thus somewhat artificial as performance measures. However, generalizations are clearly possible because of the very high quality, complexity, and diversity of the tutor requirements. Indeed, the resulting tutor score distributions that reflect performance competence and proficiency are exceptional and are about as good as one could hope for.

In terms of real-world criterion domains, Arthur, Barrett, and Doverspike (1990) successfully examined cognitive-process measures as predictors of the frequency of accidents among transport drivers. Also, Ackerman and Kanfer (1993) demonstrated the validity of a cognitive-process-based battery for predicting air traffic controller success. The current research extends these studies and provides important information regarding the usefulness of well-designed cognitive-process measures.

A major advantage of the cognitive measures used here is that the various theoretical bases of the cognitive components and abilities within the battery are clearly defined and specified. The test battery itself is highly diverse in terms of its content, the products required, and the operations used and represents a truly cutting-edge test battery. Indeed, an initial goal of the CAM research team was to measure additional components to increase the amount of criterion variance accounted for. A better match between test components and criterion space is critical and possible. Indeed, in 1981, Sternberg noted that "the incremental value of information processing scores over psychometric scores has yet to be demonstrated for interesting external criteria" (p. 1186). We believe that this correspondence and incremental value has been achieved with the current research and that this research stands in stark contrast to this statement. Having used a construct-coverage framework rather than a strict "either–or" (that is, either information processing or general ability) approach to specifying and evaluating measures has yielded promising results. Such a framework has allowed for the possibility that additional ability variance may be assessed by the cognitive-process measures.

Indeed, it is likely that the cognitive-process measures and the traditional measures were successful in part because they both contained appreciable cognitive and intellectual content (Humphreys, 1989), which is relevant for many real-world performance and learning contexts. It is also important to return to the notions of symbolic representation and adaptability as crucial elements in the original definitions of intelligence. We believe that the measures we studied successfully and uniquely captured this critical content and, in so doing, enabled the strong correspondence between predictor and real-world-emulating criterion domains. Indeed, a seemingly effortless movement into new domains on the part of individuals appears to underlie the construct space elements specified by early theory, as well as defining the measures and criteria measured in this study.

A critical part of many definitions of intelligence is the ability to acquire and use information and conceptual skills in new contexts. It is also true that reading comprehension and arithmetic reasoning are typically excellent measures of the general factor. These were represented by the V-M first-order factor, as we described earlier. Furthermore, it is important to recall that very large factor intercorrelations were observed for this factor and several of the cognitive-process first-order factors, specifically PS, WM, and IN (both score types for PS and WM). It appears, therefore, that the process measures assessing these factors are indeed excellent markers of the general factor. Also, Kyllonen and Christal (1990) found that WM and traditionally defined reasoning ability are correlated in the .8–.9 range. However, additional content, process, and operation requirements were measured with the remaining assessments. As Humphreys (1989) stated, the "ability of humans to deal with verbal, mathematical, and figural content in complex ways is a fundamental characteristic of measures having high loadings on the general factor in intelligence" (p. 322). Thus, it appears that the combined batteries and therefore the construct space underlying performance on the batteries was represented not only by the traditional elements but by novel elements, all pointing to the general factor.

Another point about the general factor is worth mentioning. As Lubinski and Dawis (1992) stated, because we live in a rapidly changing society, "familiar tasks change as they take on unfamiliar dimensions" (p. 2). They pointed out that novel, sophisticated behaviors must now be quickly learned and mastered. It may well be that these developments and realities have increased the opportunities for the manifestation of ability differences in performance on such jobs. It is highly likely that both our criteria and predictors were highly "*g*-saturated," perhaps even more so than previous studies have allowed. Such *g*-saturation may have translated into a

picture of general intelligence that is even broader and more diverse than researchers have previously acknowledged.

The results of this research are both encouraging and compelling, but a few caveats concerning limitations are warranted. One limitation is the examinee population. The question of motivation arises, as it would in any research involving nonapplicant samples completing nonoperational measures. The participants were encouraged to perform to the best of their ability, but the degree to which their motivation would have differed from that of an applicant sample is uncertain. Empirical research investigating the nature and measurement of participants' motivational states is obviously necessary (Arvey, Strickland, Drauden, & Martin, 1990). However, given the results reported earlier on the lack of range restriction on the general factor in the studied groups, along with the generally very high level of predictive validities, we can be fairly sure that motivation levels were quite high. Furthermore, as stated in the *Principles for the Validation and Use of Personnel Procedures* (Society for Industrial and Organizational Psychology, 1987), validation efforts should use samples that are "reasonably representative of the populations of people and jobs to which the results are to be generalized" (p. 7). We believe there was significant individual-differences-relevant diversity in the sample that translated into ability and performance differences. Furthermore, given that selection (direct, indirect, or self-selection; see Roznowski, 1998) and concomitant range restriction were not an issue here, generalizability to typical applicant populations can be made.

Additionally, careful psychometric work may very likely further improve such batteries as the one studied here. Indeed, standard paper-based batteries have benefited from years of close measurement. For instance, item analyses performed on the current battery revealed that the WM tests that performed very well overall had excellent item difficulty distributions. Results also indicate that the useful speed-of-processing variance can likely be assessed with a smaller number of tests overall. On the other hand, part of the problem with the latency scores may have been psychometric in nature. We know considerably less about distributions of latencies for individuals than we do about error rate measures, at least from a psychometric, individual-differences perspective. Furthermore, although acceptable coefficient alphas were found for the latency measures, other research has shown that latency measures typically have less-than-desirable stabilities (test–retest reliabilities; e.g., Roznowski, 1993). Indeed, as Stauffer, Ree, and Carretta (1996) have stated before, such "measures can be expected to be valid predictors of occupational criteria in proportion to their reliable measurement of '*g*' " (p. 199). We now know that these measures assess *g* but hope that even greater reliability and construct coverage can indeed be achieved in future batteries.

References

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

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

Allport, G. W. (1937). *Personality: A psychological interpretation.* New York: Holt, Rinehart and Winston.

Allport, G. W. (1966). Traits revisited. *American Psychologist, 21,* 1–10.

Anderson, J. R. (1983). *The architecture of cognition.* Mahwah, NJ: Erlbaum.

Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985, April). Intelligent tutoring systems. *Science, 228,* 456–462.

Arthur, W., Barrett, G. V., & Doverspike, D. (1990). The validity of an information-processing-based test battery for the prediction of handling accidents among petroleum-product transport drivers. *Journal of Applied Psychology, 75,* 621–628.

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test-taking. *Personnel Psychology, 43,* 695–716.

Baddeley, A. (1986). *Working memory.* Oxford, England: Clarendon Press.

Boring, E. G. (1923, June 6). Intelligence as the tests test it. *The New Republic,* 35–37.

Boring, E. G. (1950). *A history of experimental psychology* (Rev. ed.). New York: Appleton-Century-Crofts.

Brody, N. (1992). *Intelligence.* San Diego, CA: Academic Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Lang (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Campbell, J. P. (1990). An overview of the Army selection and classification project (Project A). *Personnel Psychology, 43,* 231–239.

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

Cattin, P. (1980). Estimation of the predictive power of a regression model. *Journal of Applied Psychology, 65,* 407–414.

Christal, R. E. (1991). *Comparative validities of ASVAB and LAMP tests for logic gates learning* (Report No. AL-TP-1991-0031). Brooks Air Force Base, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. Linn (Ed.), *Intelligence: Measurement, theory, and public policy. Proceedings of a symposium in honor of Lloyd G. Humphreys* (pp. 147–171). Urbana, IL: University of Illinois Press.

Detterman, D. K. (1982). Does "*g*" exist? *Intelligence, 6,* 99–108.

Dunnette, M. D. (1976). Aptitudes, abilities, and skills. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (1st ed., pp. 473–520). Chicago: Rand McNally.

Embretson, S. E. (1983). Construct validity: Construct representation and nomothetic span. *Psychological Bulletin, 93,* 179–197.

Fiske, D. W., & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin, 112,* 393–395.

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Guilford, J. P. (1985). The structure of intellect model. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications* (pp. 225–266). New York: Wiley.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning and discovery.* Cambridge, MA: MIT Press.

Humphreys, L. G. (1971). Theory of intelligence. In R. Cancro (Ed.), *Intelligence: Genetic and environmental influences* (pp. 31–42). New York: Grune & Stratton.

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

Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and learning* (pp. 87–120). New York: Plenum.

Humphreys, L. G. (1984). General intelligence. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives in mental testing* (pp. 221–248). New York: Plenum.

Humphreys, L. G. (1985). Attenuated hypothesis or attenuated test of hypothesis? *Intelligence, 9*(3), 291–295.

Humphreys, L. G. (1989). The first factor extracted is an unreliable

estimate of Spearman's "*g*": The case of discrimination reaction time. *Intelligence, 13,* 319–324.

Humphreys, L. G., & Parsons, C. K. (1979). Piagetian tasks measure intelligence and intelligence tests assess cognitive development. *Intelligence, 3,* 369–382.

Hunt, E., & Pellegrino, J. W. (1985). Using interactive computing to expand intelligence testing. *Intelligence, 9,* 207–236.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29,* 340–362.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternate predictors of job performance. *Psychological Bulletin, 96,* 72–98.

Irvine, S. H., Anderson, J. D., & Dann, P. L. (1990). Towards a theory of algorithm-determined cognitive test construction. *British Journal of Psychology, 81,* 173–195.

Jensen, A. R. (1998). *The general factor in human intelligence.* New York: Praeger.

Kyllonen, P. C. (1991). Principles for creating a computerized test battery. *Intelligence, 15,* 1–15.

Kyllonen, P. C. (1994). Cognitive abilities testing: An agenda for the 1990s. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 103–125). Hillsdale, NJ: Erlbaum.

Kyllonen, P. C., & Christal, R. E. (1989). Cognitive modeling of learning abilities: A status report of LAMP. In R. F. Dillon & J. W. Pellegrino (Eds.), *Testing: Theoretical and applied perspectives.* (pp. 146–173). New York: Praeger.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working memory capacity?! *Intelligence, 14,* 389–433.

Kyllonen, P. C., Woltz, D. J., Christal, R. E., Tirre, W. C., Shute, V. J., & Chaiken, S. (1990). *CAM-4: Computerized battery of cognitive ability tests.* Unpublished computer program, Brooks Air Force Base, TX.

Lesgold, A., Busszo, M. S., McGinnis, T., & Eastman, R. (1989). *Windy, the flight engineering tutor* [Unpublished computer software]. Learning Research and Development Center, University of Pittsburgh, PA.

Loftus, E. F. (1983). Activation of semantic memory. *American Journal of Psychology, 86,* 331–337.

Lubinski, D., & Dawis, R. V. (1992). Aptitudes, skills, and proficiencies. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3, pp. 1–59). Palo Alto, CA: Consulting Psychologists Press.

Lubinski, D., & Humphreys, L. G. (1990). A broadly based analysis of mathematical giftedness. *Intelligence, 14,* 327–355.

Lubinski, D., & Humphreys, L. G. (1992). Some bodily and medical correlates of mathematical giftedness and commensurate levels of socioeconomic status. *Intelligence, 16,* 99–115.

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. *Psychological Bulletin, 97,* 562–582.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114,* 449–458.

Meehl, P. E. (1986). Trait language and behaviorese. In T. Thompson & M. Zeiler (Eds.), *Analysis and integration of behavioral units* (pp. 335–354). Hillsdale, NJ: Erlbaum.

Meehl, P. E., & Cronbach, L. J. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Pintner, R. (1921). Contribution to "Intelligence and its measurement." *Journal of Educational Psychology, 12,* 139–143.

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

Ree, M., & Earles, J. (1991). Predicting training success: Not much more than *g*. *Personnel Psychology, 44,* 321–332.

Ree, M., & Earles, J. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science, 1,* 86–89.

Roznowski, M. (1993). Measures of cognitive processes: Their stability and other psychometric and measurement properties. *Intelligence, 17,* 361–388.

Roznowski, M. (1998). The Graduate Record Examination: Testing:: Yale: Academics. *American Psychologist, 53,* 570–572.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Shute, V. J. (1991). Who is likely to acquire programming skills? *Journal of Educational Computing Research, 7,* 1–24.

Shute, V. J. (1993a). A comparison of learning environments: All that glitters. . . . In S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 47–73). Hillsdale, NJ: Erlbaum.

Shute, V. J. (1993b). A macroadaptive approach to tutoring. *Journal of Artificial Intelligence and Education, 4,* 61–93.

Shute, V. J., & Psotka, J. (1996). Intelligent tutoring systems: Past, present, and future. In D. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 570–600). New York: Macmillan.

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.

Society for Industrial and Organizational Psychology. (1987). *Principles for the validation and use of personnel selection procedures: Third edition.* College Park, MD: Author.

Spearman, C. (1904). "General intelligence": Objectively determined and measured. *American Journal of Psychology, 15,* 201–292.

Spearman, C. (1923). *The nature of intelligence and the principles of cognition.* London: Macmillan.

Spearman, C. (1927). *Abilities of man: Their nature and measurement.* New York: MacMillan.

Stauffer, J. M., Ree, M. J., & Carretta, T. R. (1996). Cognitive-components tests are not much more than *g*: An extension of Kyllonen's analyses. *The Journal of General Psychology, 23,* 193–205.

Sternberg, R. J. (1981). Testing and cognitive psychology. *American Psychology, 36,* 1181–1189.

Vernon, P. E. (1969). *Intelligence and cultural environment.* London: Methuen.

Wigdor, A. K., & Green, B. F., Jr. (1991). *Performance assessment in the workplace.* Washington, DC: National Academy Press.

Woltz, D. J. (1988). An investigation of the role of working memory in procedural skill acquisition. *Journal of Experimental Psychology: General, 117,* 319–331.

Received June 23, 1997
Revision received January 10, 2000
Accepted January 13, 2000 ■