INTELLIGENCE **20,** 15–25 (1995)

Validity Results for *g* From an Expanded Test Base

GERALD E. LARSON
JOHN H. WOLFE
Navy Personnel Research and Development Center, San Diego, CA

When vocational aptitude batteries are expanded by adding new tests, the most common way to measure validity gains is to regress various criteria onto the subtest scores from the old and new batteries and contrast the results. A rarely tried approach that may be of equal value, however, is to examine "accidental" validity gains for a recalculated general intelligence (or *g*) score based on the new battery, because many psychologists have argued that the majority of test validity comes from *g* rather than specific abilities. In this article, we examine validity differences for a *g* score calculated on the Armed Services Vocational Aptitude Battery (ASVAB) alone versus the ASVAB plus nine diverse experimental tests selected for their potential importance and uniqueness from the ASVAB. Although no validity gain for expanded *g* was observed for final school grade criteria, a 6% validity gain was obtained for hands-on performance measures. A gain of this size is of practical importance in the armed forces.

When vocational aptitude batteries are expanded by adding new tests, the most common way to measure validity gains is to regress various criteria onto the subtest scores from the old and new batteries and contrast the results. Validity gains are expected to result from specific relationships between the just-added constructs and one or more criterion scores. For example, one may find that the maximum gains from adding a spatial test accrue in mechanical jobs where task analyses reveal many spatial-type job demands. This, at least, is the idealized scenario portrayed in the ability profile view, or, as Jensen (1984) refers to it, the "specificity doctrine." The doctrine holds that mental abilities consist of a repertoire of specific skills and that test batteries measure some selected sample of the skill repertoire. The broader the sampling, the more types of outcomes may be successfully predicted.

A rarely tried approach that may be of equal value, however, is to examine

The authors would like to thank David Alderton for his invaluable role in the ECAT project. We would also like to thank Norm Abrams of RGI Inc. for suggesting the comparison of academic and performance measures as part of the validity analyses herein. The opinions expressed are those of the authors, are not official, and do not necessarily reflect the views of the Department of the Navy.

Correspondence and requests for reprints should be sent to Gerald E. Larson, Personnel Systems Department, Navy Personnel Research and Development Center, 53335 Ryne Road, San Diego, CA 92152–7250.

"accidental" validity gains for a recalculated general intelligence (or *g*) score based on the new battery, because many psychologists (e.g., Humphreys, 1979; Hunter, 1986; Jensen, 1984; McNemar, 1964; Ree & Earles, 1991; Thorndike, 1985) have argued that the majority of test validity comes from *g* rather than specific abilities. Hunter (1986), for example, noted that the massive databases gathered by the U.S. Employment Service and the Armed Forces clearly suggest that it is general cognitive ability rather than specific cognitive aptitudes that predicts job performance. Moreover, there are reasons for believing that some, but not all, test battery changes could provide a theoretically better (and perhaps more valid) *g* estimate. First of all, if a battery is narrowly based to begin with, then diversification should reduce specificity in a factor representing the largest pool of common variance. For example, contrast the positive manifold among a set of verbal tests with the positive manifold among a collection of verbal and spatial tests. Only the latter approaches what we typically mean by *general* intelligence. Also, if a battery lacks a fluid intelligence test (as some scholastic aptitude batteries do, for example) then adding a fluid test may improve the measurement of *g* simply by virtue of the extremely close relationship between the two constructs (Marshalek, Lohman, & Snow, 1983; Undheim & Gustafsson, 1987).

Fortunately we are now in the position of being able to test these hypotheses about accidental validity gains for *g*. The Armed Forces are concluding a multiyear effort to improve the validity of the Armed Services Vocational Aptitude Battery (ASVAB). The approach has been to broaden the range of skills measured and thereby provide an expanded ability profile with which to match people to jobs. The main analyses involved multiple regression, where the various subtest scores were allowed to enter freely into the predictive models for various school performance criteria. Those results are reported in Wolfe, Alderton, and Larson (1994). In this article, however, we examine validity differences for a *g* score calculated on the ASVAB alone versus the ASVAB plus nine diverse experimental tests selected for their potential importance and uniqueness from the ASVAB. Thus, it is by design that the experimental tests are largely nonverbal and process based, in contrast to the ASVAB, which is heavily verbal and knowledge based. The experimental battery is called ECAT (for Enhanced Computer-Administered Tests). The ASVAB and ECAT batteries are described later.

METHOD

Subjects

As part of a study to assess the validity of military aptitude tests (see Wolfe et al., 1994), data were gathered on over 11,700 enlisted military personnel in the Navy, Army, and Air Force. In this article, we focus on a subgroup of 3,922 of these individuals for whom both academic and performance-based criterion

scores were available, although in some cases data for the larger sample are also cited. Individuals were tested on the ECAT battery (described later) during basic training or "boot camp," prior to entering more specialized training in one of the occupational schools that provided criterion measures for the study. Armed Services Vocational Aptitude Battery scores were gathered from the recruits' personnel records. The sample was 95.5% male, and 97.5% used English as their dominant language. The average subject was approximately 19 years old. Nearly 84% of the sample had obtained a high school diploma, and additional 6.7% had at least some college-level schooling; only 9.5% failed to complete high school. For descriptive purposes the subjects were divided into six ethnic groups: White (71.1%), Black (16.5%), Hispanic (5.9%), Asian (2.2%), North American/Indian (0.8%), and Other/Unknown (3.4%).

Aptitude Tests

ASVAB Content. The ASVAB is a set of 10 tests used for selection and classification of military applicants. Table 1 shows the ASVAB tests and constructs.

TABLE 1
ASVAB Tests and Constructs

Construct	Test	Description
Verbal Ability	Paragraph Comprehension (PC)	A 15-item reading comprehension test.
	Word Knowledge (WK)	A 35-item vocabulary test using synonyms or words embedded in sentences.
	General Science (GS)	A 25-item knowledge test of physical and biological sciences.
Math Ability	Arithmetic Reasoning (AR)	A 30-item arithmetic word problem test.
	Math Knowledge (MK)	A 25-item test of algebra, geometry, fractions, decimals, and exponents.
Technical Knowledge	Mechanical Comprehension (MC)	A 25-item test of mechanical and physical principles.
	Auto and Shop Information (AS)	A 25-item knowledge test of automobiles, shop practices, tools, and tool use.
	Electronics Information (EI)	A 20-item test about electronics, radio, and electrical principles and information.
Clerical Skills	Numerical Operations (NO)	A 50-item speeded addition, subtraction, multiplication, and division test using one- or two-digit numbers.
	Coding Speed (CS)	An 84-item speeded test requiring the recognition of number strings arbitrarily associated with words in a table.

The tests represent Verbal Ability, Mathematical Ability, Technical Knowledge, and Clerical Ability.

ECAT Content. The goal of the (approximately) 3-hr ECAT was to broaden the ASVAB. Table 2 shows the ECAT tests and constructs. The tests represent Nonverbal Reasoning, Spatial Ability, Psychomotor Skill, and Perceptual Speed. Further details concerning the ECAT tests are presented in Wolfe et al. (1994). The battery was presented on Hewlett-Packard Integral microcomputers operating under UNIX™. Tests 1 through 6 in Table 2 used a simplified keyboard. The keyboard was modified by using a plastic mask that revealed only the designated response keys along with a key labeled *HELP* that could be pressed during testing to suspend the program and request assistance. The *S, F, H, K,* and *;* keys were relabeled as *A, B, C, D,* and *E*. The space bar was relabeled *ENTER*. The numeric keypad keys retained their meanings. Tests 7 through 9 (One-Hand Tracking, Two-Hand Tracking, and Target Identification) used a custom built "response pedestal" with response buttons, sliders, and a joystick. Two test administration sequences for ECAT were used, corresponding to odd and even social security numbers.

TABLE 2
ECAT Tests and Constructs

Construct	Test	Description
Nonverbal Reasoning	Mental Counters (CT)	A 40-item working memory test using figural content; a nonverbal reasoning test.
	Sequential Memory (SM)	A 35-item working memory test using numerical content; a nonverbal reasoning test.
	Figural Reasoning (FR)	A 35-item series extrapolation test using figural content; a nonverbal reasoning test
Spatial Ability	Integrating Details (ID)	A 40-item spatial problem-solving test.
	Assembling Objects (AO)	A 32-item spatial and semimechanical test.
	Spatial Orientation (SO)	A 24-item spatial apperception/rotation test.
Psychomotor Skill	One-Hand Tracking (T1)	An 18-item single limb psychomotor tracking test.
	Two-Hand Tracking (T2)	An 18-item multilimb psychomotor tracking test.
Perceptual Speed	Target Identification (TI)	A 36-item RT-based figural perceptual speed test.

TABLE 3
Schools Providing Criterion Measures for the Study

Training School	N	Extra Performance Measure
Field Artillery Fire Support Specialist	821	Firing composite
Apprentice Personnel Specialist	446	Words per minute typing
Air Traffic Control (Air Force)	484	Basic approach control operation
Air Traffic Control (Navy)	72	Mean of four performance tests
Aviation Electrician's Mate	278	Average of performance tests
Aviation Structural Mechanic	244	Average of performance and practical work
Aviation Ordnanceman	234	Average of all practical work
Avionics Technician	544	Average of all performance tests
Electronics Technician—Advanced	86	Average of performance tests
Operations Specialist	713	Average of all performance tests

Criterion Measures

Criterion measures in the study consisted of final school grades (FSGs) and additional hands-on type performance criteria (such as shop, laboratory, and simulator tasks) from 10 military training schools. In many cases the performance measures represent composites of multiple exercises. Development and reliability of performance composites is described in detail by Kieckhaefer et al. (1992) and summarized in Wolfe et al. (1994). Table 3 shows the schools in the study, along with the additional performance measures obtained from each school. The choice of these particular schools for study was based on two considerations: large sample size and representativeness. Attempts were made to sample schools that were exemplars of an entire job category.

RESULTS

Descriptive statistics for the ASVAB and ECAT tests (for the full sample of 11,700 subjects) are shown in Table 4. The ASVAB tests are scaled to a mean of 50 and a standard deviation of 10 in an unselected, nationally representative sample. The first six ECAT scores represent percentage correct, the two tracking scores represent tracking error in screen pixel units, and Target Identification produces an RT score.

To calculate measures of g for the test batteries, the LISREL hierarchical, confirmatory factor analysis procedure was applied to the test data.[1] The input data for the LISREL procedure were test means, standard deviations, and intercorrelations (corrected for range restriction). In specifying the model for the first-order factor structure for ASVAB and ASVAB + ECAT, factor structures empiri-

[1]All LISREL analyses were performed under contract by Fritz Drasgow.

TABLE 4
Descriptive Statistics for Aptitude Tests

Tests	*M*	*SD*
ASVAB		
Paragraph Comprehension	53.23	5.74
Word Knowledge	53.04	5.35
General Science	53.26	7.42
Arithmetic Reasoning	53.61	6.91
Math Knowledge	55.13	6.88
Mechanical Comprehension	54.99	7.70
Auto and Shop Information	53.61	8.05
Electronics Information	52.59	7.95
Numerical Operations	54.21	6.58
Coding Speed	53.25	6.94
ECAT		
Mental Counters	0.72	0.18
Sequential Memory	0.69	0.13
Figural Reasoning	0.67	0.19
Integrating Details	0.76	0.13
Assembling Objects	0.63	0.19
Spatial Orientation	0.52	0.28
One-Hand Tracking	2,765	392
Two-Hand Tracking	3,639	472
Target Identification	1.835	0.61

cally derived by Alderton and Larson (1992) were used. Because the purpose of the study was to calculate validities for different representations of *g*, factor scores had to be created. Because LISREL does not automatically produce factor score coefficients, the coefficients were calculated using regression procedures (details of which are available from the authors upon request). The *g*-factor loadings and factor-score coefficients for the ASVAB and ASVAB + ECAT batteries are shown in Table 5. Once factor scores were calculated, the range-corrected correlation between the two *g* measures was found to be .92. To allow comparison between the ASVAB and ECAT batteries, similar procedures were used to calculate *g* for ECAT alone. The range-corrected correlation between ASVAB *g* and ECAT *g* was .73. Finally, we calculated the congruence coefficient between the two vectors of ASVAB *g* loadings (10 per vector) shown in Table 5. The congruence coefficient was an extremely high .998, indicating that *g* loadings remain highly stable even when a battery is nearly doubled in size and the newly introduced tests are quite different from the original tests.

Correlations Between *g* and School Performance

To correct the criterion means and standard deviations for range restriction, Lawley's (1943) multivariate range correction procedure was used, with all 10 AS-

TABLE 5
***g* Loadings *(A)* and *g* Factor Score Coefficients *(B)* for the Two Batteries**

	ASVAB		ASVAB + ECAT	
Test	***A***	***B***	***A***	***B***
Paragraph Comprehension	.6928	.1084	.6059	.0402
Word Knowledge	.7548	.1836	.6702	.0769
General Science	.7239	.1020	.6786	.0421
Arithmetic Reasoning	.8353	.3565	.8064	.1876
Math Knowledge	.7218	.1451	.7177	.0926
Mechanical Comprehension	.6703	.0925	.7120	.0820
Auto and Shop Information	.5032	.0259	.5043	.0164
Electronics Information	.5423	.0391	.5429	.0248
Numerical Operations	.5065	.0789	.4625	.0392
Coding Speed	.4521	.0467	.4367	.0301
Mental Counters			.7023	.0986
Sequential Memory			.6701	.0838
Figural Reasoning			.7151	.1068
Integrating Details			.7363	.1109
Assembling Objects			.7187	.1026
Spatial Orientation			.6864	.0827
One-Hand Tracking			−.4653	−.0244
Two-Hand Tracking			−.5144	−.0495
Target Identification			−.4180	−.0488

VAB tests used as explicitly selected variables. The unrestricted reference population was the 1991 set of applicants to the armed services (N = 650,278). The corrected reliability was computed following Gulliksen (1950/1987). Corrected reliabilities were used for correcting validities. The top half of Table 6 shows the corrected validities for ASVAB *g* and ASVAB + ECAT *g*, with final school grades serving as criterion measures.

As the table shows, the high intercorrelation between the two *g* measures also corresponds to average validities that are identical (.77). This validity is considerably higher than the mean validity of .61 reported for ASVAB *g* by Ree and Earles (1991) in their study of 78,041 Air Force enlistees in 82 training schools, and it is also higher than the correlation of .63 between general cognitive ability and military job (vs. training) performance reported in a large-scale Army study (McHenry, Hough, Toquam, Hanson, & Ashworth, 1990). The difference between our results and those of Ree and Earles might either be due to our correction for criterion unreliability (a step not performed by Ree and Earles), or to the fact that many of the jobs sampled for the current study are highly technical in nature, perhaps leading to greater validities for *g*. The latter point may also help to explain the difference between our findings and those of McHenry et al. (1990), whose subjects included a large number of infantry and tank crew mem-

TABLE 6
Validity of Two *g* Measures in Schools With FSG and Performance Criteria

Training School	Correlations With FSG	
	ASVAB *g*	(ASVAB + ECAT) *g*
Field Artillery Fire Support Specialist	.77	.80
Apprentice Personnel Specialist	.83	.80
Air Traffic Control (Air Force)	.72	.72
Air Traffic Control (Navy)	.79	.77
Aviation Electrician's Mate	.66	.68
Aviation Structural Mechanic	.81	.78
Aviation Ordnanceman	.70	.69
Avionics Technician	.78	.78
Electronics Technician—Advanced	.81	.84
Operations Specialist	.78	.78
Weighted Average	.77	.77

Training School	Correlations With Performance Criteria	
	ASVAB *g*	(ASVAB + ECAT) *g*
Field Artillery Fire Support Specialist	.71	.71
Apprentice Personnel Specialist	.33	.34
Air Traffic Control (Air Force) A[a]	.51	.63
Air Traffic Control (Air Force) B[a]	.41	.51
Air Traffic Control (Navy)	.31	.43
Aviation Electrician's Mate	.58	.61
Aviation Structural Mechanic	.57	.60
Aviation Ordnanceman	.45	.48
Avionics Technician	.54	.60
Electronics Technician—Advanced	.68	.71
Operations Specialist	.75	.79
Weighted Average	.58	.62

[a]Air Traffic Control students were split into two groups following a curriculum change midway through the study.

bers. In any event, all the studies indicate that the typical validity of *g* in military settings is at least .60 and may be even higher under certain conditions.

As noted earlier, each of the schools in the study reported student performance on a variety of hands-on performance exercises in addition reporting FSGs (based largely on written exams). Because these hands-on exercises are closely related to subsequent job demands, and because the prediction of job performance is a primary goal of aptitude testing, the validation of *g* indices against hands-on performance scores is an important endeavor. The bottom half of Table

6 shows validity results for ASVAB *g* and ASVAB + ECAT *g*, using hands-on performance scores as criteria. Results for *performance* measures indicate that, unlike the case with FSGs, the validity of ASVAB + ECAT *g* was 4 correlation points (about 6%) higher than the validity of ASVAB *g*. Moreover, the pattern is quite consistent, because in 10 out of 11 comparisons validity of ASVAB + ECAT *g* is higher than that of ASVAB *g*. The practical implications of this finding are discussed in the following.

DISCUSSION

Given the considerable evidence for the importance of general mental ability, it is critical to determine the best test combination with which to assess *g*. At issue is whether all multidimensional aptitude batteries provide equally valid estimates of *g*, or whether certain types of test batteries have unique advantages. The first view has much support. According to Brody (1992), Charles Spearman (the "discoverer" of *g*) believed that the aggregate *g* score obtained on one set of diverse subtests would be in substantial agreement with the aggregate score obtained on a different battery of diverse tests, an argument known as the principle of the *indifference of the indicator*. A similar point is emphasized by Jensen (1993), who argued that the *g* factor does not fluctuate capriciously from one collection of tests to another. For example, Jensen (1993) indicated that even though the six Verbal subtests of the Wechsler Intelligence Scale for Children (WISC) look very different from the six Performance subtests, the *g* extracted from just the Verbal subtests is correlated about .80 with the *g* extracted from just the Performance subtests.

On the other hand, it is known that tests vary in their individual relationships with *g*, and that fluid intelligence tests in particular have strong *g* relationships. For example, when a general intelligence factor is extracted under what are probably optimal conditions, that is, when highly diverse mental tests are administered to examinees ranging broadly in ability, the *g* factor is virtually indistinguishable from a fluid ability factor (Marshalek et al., 1983; Undheim & Gustafsson, 1987). Therefore, a test battery with a number of fluid ability tests may provide a more construct valid measure of *g* than would be obtainable on a battery lacking fluid ability subtests. Moreover, differences in predictive validity might also be observed.

Our results indicate that conclusions about validity gains depend partly on the nature of the criterion measure. For scores on written exams (summarized in the FSG measures), the validity of *g* remained constant as more tests were added to the battery, supporting the principle of the indifference of the indicator. For practical or hands-on type measures, however, a 6% validity gain was observed, despite the fact that the intercorrelation between the two *g* measures was .92. Although it remains unclear why results were criterion-dependent, two explanations are worth consideration. First, the average correlation between ASVAB *g*

and FSG was an exceptionally high .77, leaving much less room for improvement than was the case with performance criteria where the average validity for ASVAB *g* was a more modest .58. Thus, it may simply be the case that the validity of any single test score across diverse occupations is unlikely ever to exceed some value of about .75. Indeed, few validities higher than .70 seem to exist in the literature. An alternative explanation for the criterion-dependent results would hold that specific job content in the armed forces overlaps more with test content in the ASVAB + ECAT batteries than with the ASVAB battery alone. However, because *g* scores represent general rather than specific test variance, this latter explanation seems somewhat suspect.

Though a 6% validity gain for predicting performance criteria may seem modest, it is important to note that even slight validity gains can have substantial economic benefits when applied to large organizations such as the armed forces. For example, Schmidt, Hunter, and Dunn (1987) estimated that an increase in ASVAB validity of 3% would result in the equivalent of $83 million annually in performance improvement in the Navy. The savings occur because the additional validity would result in more effective assignment of personnel. In summary, although most validation efforts are focused on detecting relationships between specific subtest scores and various outcome measures, our results suggest that one should also measure validity changes for *g* following changes to test batteries, particularly because most test validity is thought to come from *g* (e.g., Humphreys, 1979; Hunter, 1986; Jensen, 1984; McNemar, 1964; Ree & Earles, 1991; Thorndike, 1985).

REFERENCES

Alderton, D.L., & Larson, G.E. (1992, November). *Navy and joint-service validity research: Interim conclusions.* Paper presented to the NATO Research Group 15 on Computer-Based Selection and Classification in the Military, Monterey, CA.

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

Gulliksen, H. (1987). *Theory of mental tests.* Hillsdale, NJ: Erlbaum. (Original work published 1950)

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

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

Jensen, A.R. (1984). Test validity: *g* versus the specificity doctrine. *Journal of Social and Biological Structures, 7,* 93–118.

Jensen, A.R. (1993). Spearman's *g*: Links between psychometrics and biology. *Annals of the New York Academy of Sciences, 107,* 103–129.

Kieckhaefer, W.F., Ward, D.G., Kusulas, J.W., Cole, D.R., Rupp, L.M., & May, M.H. (1992). *Criterion development for 18 technical training schools in the Navy* (Contract #N66001-90-D-9502, Delivery Order 7J08). San Diego, CA: Navy Personnel Research and Development Center.

Lawley, D. (1943). A note on Karl Pearson's selection formulae. *Royal Society of Edinburgh Proceedings, Section A, 62,* B28–30.

Marshalek, B., Lohman, D.F., & Snow, R.E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence, 7,* 107–127.

McHenry, J.J., Hough, L.M., Toquam, J.L., Hanson, M.A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43,* 335–354.

McNemar, Q. (1964). Lost: Our intelligence? Why? *American Psychologist, 19,* 871–882.

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

Schmidt, F.L., Hunter, J.E., & Dunn, W.L. (1987). *Potential utility increases from adding new tests to the Armed Services Vocational Aptitude Battery (ASVAB).* Unpublished manuscript, Navy Personnel Research and Development Center, San Diego, CA.

Thorndike, R.L. (1985). The central role of general ability in prediction. *Multivariate Behavior Research, 20,* 241–254.

Undheim, J.O., & Gustafsson, J.E. (1987). The hierarchical organization of cognitive abilities: Restoring general intelligence through the use of linear structural relations (LISREL). *Multivariate Behavioral Research, 22,* 149–171.

Wolfe, J.H., Alderton, D.L., & Larson, G.E. (1994). *Incremental validity of Enhanced Computer Administered Testing (ECAT).* Manuscript submitted for publication.