A Comparison of Item Calibration Media in Computerized Adaptive Testing

Rebecca D. Hetter and Daniel O. Segall
Navy Personnel Research and Development Center
Bruce M. Bloxom
Department of Defense Manpower Data Center

A concern in computerized adaptive testing is whether data for calibrating items can be collected from either a paper-and-pencil (P&P) or a computer administration of the items. Fixed blocks of power test items were administered by computer to one group of examinees and by P&P to a second group. These data were used to obtain computer-based and P&P-based three-parameter logistic model parameters of the items. Then each set of parameters was used to estimate item response theory pseudo-adaptive scores for a third group of examinees who had received all of the items by computer. The effect of medium of administration of the calibration items was assessed by comparative analyses of the adaptive scores using structural modeling. The results support the use of item parameters calibrated from either P&P or computer administrations for use in computerized adaptive power tests. The calibration medium did not appear to alter the constructs measured by the adaptive test or the reliability of the adaptive test scores. *Index terms: computerized adaptive testing, item calibration, item parameter estimation, item response theory, medium of administration, trait level estimation.*

Computerized adaptive tests (CATs) provide efficient assessment of psychological constructs (see Weiss, 1983). When combined with item response theory (IRT), CATs use item parameter estimates to select the most informative item for administration at each stage of the assessment; these parameters also are used to update both point and interval estimates of the examinee's score.

A practical concern in the initial development of a CAT is whether the items must be calibrated from data collected in a computerized administration or whether equally accurate results can be obtained by calibrating the items from data collected in a paper-and-pencil (P&P) administration. For example, in the development of the CAT version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB), item parameter estimates were available only from a P&P administration of the items (Prestwood, Vale, Massey, & Welsh, 1985) because computers were not available at the testing sites. This made it important to assess whether scores obtained on the CAT-ASVAB using the P&P-based item calibration had the same interpretation and precision as scores obtained from a computer-based calibration of the items.

Previous research comparing the effects of computer-based and P&P-based administration of cognitive tests has been concerned primarily with the medium or mode of administration (MOA) of the actual test rather than the MOA used for item calibration. Although this research has not always explicitly addressed the effects on CATs, it has provided a number of results that are suggestive of the potential importance of three effects of the MOA.

Previous studies have examined the effect of the MOA on the construct assessed by the test. Observed-score factor analytic and correlational studies by Moreno, Wetzel, McBride, & Weiss (1984) and Moreno, Segall, & Kieckhaefer (1985) suggested that the factor pattern of a cognitive battery has the same hyperplane pattern whether the tests are administered by conventional P&P or adaptively by computer. Also, a meta-analytic study by Mead &

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 18, No. 3, September 1994, pp. 197–204

0146-6216/94/030197-08$1.65

Drasgow (1993) indicated that correlations close to 1.00 were obtained between computerized and P&P versions of the same test when the correlations were corrected for attenuation and the tests were power tests; this result was found whether the computerized tests were adaptive or nonadaptive. The findings of Mead and Drasgow imply that the disattenuated correlations among tests of different traits were essentially the same whether the traits were measured using the same MOA or a different MOA. However, this implication has yet to be tested empirically.

Previous studies also have examined the effect of the MOA on the precision of the test. Green, Bock, Linn, Lord, & Reckase (1984) suggested that nonsystematic effects of the MOA could degrade the precision of CATs if the tests were administered and scored using P&P-based item calibrations. They noted that such effects could arise when some items were affected (e.g., in difficulty) by the MOA and other items were not. Divgi (1986) and Divgi & Stoloff (1986) studied this problem empirically and found that item response functions (IRFs) estimated from items administered adaptively by computer differed from IRFs obtained from a conventional P&P administration of the same items. However, these differences were found not to be systematically related to the content of the items and, when applied to the scoring of adaptively administered items, produced only slight effects on final test scores. Also, Moreno & Segall (1992) showed that even if nonsystematic effects of calibration error result from using a P&P-based calibration in an adaptive test, the adaptive test still can have greater reliability than a longer, conventional P&P test. Although these results are reassuring about the relative precision of CATs and conventional P&P tests, what remains to be demonstrated is whether the medium used to obtain data for computing item parameters affects the precision of a CAT; specifically, whether nonadaptively computer-administered items produce a calibration that results in CAT scores with greater reliability than scores produced from a P&P-based calibration.

Previous studies also have investigated the effect of the MOA on the score scale of the test. Green et al. (1984) suggested that the MOA could have a systematic effect on the score scale; for example, by making all of the items on the test more difficult or easier to a similar extent. Empirical results reported by Spray, Ackerman, Reckase, & Carlson (1989) and meta-analyzed by Mead & Drasgow (1993) indicated that computer-administered items can result in slightly lower mean cognitive test scores than P&P-administered items. Spray et al. (1989) also used the regressions of items on total scores to investigate whether the effects were general to all items or specific to some items; however, they found no MOA effect for most of their items, which made their results inconclusive. What remains to be investigated is whether the effects of the MOA on the score scale of a test are systematic—that is, removable by a transformation (e.g., linear) of the score scale—or nonsystematic—that is, altering the reliability of scores by affecting the difficulty of some items but not others.

Purpose

The primary purpose of this study was to compare the effects on CAT scores of using a P&P calibration versus a computer calibration to select items and estimate scores. The two primary effects under investigation were the effects on (1) the construct being assessed and (2) the reliability of the scores. The specific question for each effect was the extent to which adaptive scores obtained with computer-administered items and a P&P calibration correspond to adaptive scores obtained with the same computer-administered items (and responses) and a computer calibration. A secondary effect under investigation was the influence of calibration medium on the score scale. The specific question for this effect was the extent to which IRT difficulty parameters obtained with a P&P calibration corresponded to those obtained for a calibration of the same items from a non-adaptive computer administration.

Method

At each testing session, examinees were randomly assigned to one of three groups. Fixed blocks of power test items were administered by computer to one group of examinees (Group 1) and by P&P to a second group (Group 2). These data were used to obtain computer-based and P&P-based three-parameter logistic model

(3PLM) calibrations of the items. Each calibration then was used to estimate IRT adaptive scores or trait levels (θs) for a third group of examinees who were administered the items by CAT real-data simulation (Group 3). The effect of the calibration MOA (CMOA) on the construct being assessed and on the reliability of the test scores was assessed by comparative analyses of the $\hat{\theta}$s using the alternative calibrations. The effect of the CMOA on the score scale was assessed by a comparative analysis of IRT difficulty parameters from computer-based and P&P-based calibrations.

Examinees

The examinees were Navy recruits stationed at the Recruit Training Center in San Diego. The total number of examinees was 2,955. There were 989 examinees in Group 1, 978 in Group 2, and 988 in Group 3. These sample sizes were adequate for independent calibrations. A simulation study by Hulin, Drasgow, & Parsons (1983, pp. 101–110) suggested that larger samples produce little improvement in the precision of IRFs and scores, given the number of items (40) used in these calibrations. ASVAB scores were obtained from file data for nearly all examinees and were used to assess whether the groups were comparable in ability levels.

Calibration Tests

The items were taken from power test pools specifically developed for the CAT-ASVAB by Prestwood et al. (1985). 40 items from each of four content area tests—general science (GS), arithmetic reasoning (AR), word knowledge (WK), and shop information (SI)—were used (160 items in total). Although only four of the 11 CAT-ASVAB tests were included in this study, the MOA tests were administered in the same order as in the CAT-ASVAB.

The three groups received exactly the same items with the same instructions, the same practice problems, in the same order, and with the same time limits. The items were conventionally administered in order of ascending difficulty, using the 3PLM difficulties obtained by Prestwood et al. (1985). The P&P test employed a booklet and an optically scanned answer sheet; the booklet format was the same as that used in the original P&P calibration by Prestwood et al. (1985). The computer-administration format was the same as that used in the CAT-ASVAB (one item per screen, no return to previous items, omits not allowed). Practice problems and instructions were printed on the booklet and read aloud by the proctor for the P&P group (Group 2), and presented on the screen, with the option to repeat, for the computer groups (Groups 1 and 3).

The tests were timed; however, time limits were liberal. They were prorated from 95% completion times obtained in previous research for the same tests, with the addition of 10% to allow for a higher completion rate. Test order and time limits were as follows: GS (19 min), AR (63 min), WK (16 min), and SI (17 min).

Item Calibrations

IRT parameter estimates based on the 3PLM (Birnbaum, 1968) were obtained in separate calibrations for computer Group 1 (calibration C1) and for P&P Group 2 (calibration C2). The response datasets on which the calibrations were based were labeled U1 and U2, respectively. The calibrations were performed with LOGIST 6 (Wingersky, Barton, & Lord, 1982), a computer program that uses a joint maximum likelihood approach. Response dataset U3 from Group 3 (the second computer group) was not used in the calibrations. The design with the corresponding notation is summarized in Table 1.

Table 1
Calibration Design

Group	MOA	Response Dataset	Item Parameter Calibration
1	Computer	U1	C1
2	P&P	U2	C2
3	Computer	U3	—

Scores

For each examinee in Group 3, two $\hat{\theta}$s were computed for each test (see Table 2). All $\hat{\theta}$s were based on the U3 responses. $\hat{\theta}$s for variables X_{GSC}, X_{ARC}, X_{WKC}, and X_{SIC} (where C is computer CMOA) were calculated using the computer-based item parameters (C1). Scores for variables X_{GSP}, X_{ARP}, X_{WKP}, and X_{SIP}

(where P is P&P CMOA) were calculated using the P&P-based item parameters (C2). All $\hat{\theta}$s were based on simulated CATs, computed as described below, using only 10 of the 40 responses from a given examinee.

Table 2
Variable Definitions for the Scores Computed From Response Dataset U3 for Two CMOAs

Variable	Test	CMOA
X_{GSC}	GS	Computer
X_{ARC}	AR	Computer
X_{WKC}	WK	Computer
X_{SIC}	SI	Computer
X_{GSP}	GS	P&P
X_{ARP}	AR	P&P
X_{WKP}	WK	P&P
X_{SIP}	SI	P&P

Adaptive scores. To compute the adaptive $\hat{\theta}$s, 10-item adaptive tests were simulated using actual examinee responses. As in CAT-ASVAB, a normal (0,1) prior distribution of θ was assumed, Owen's (1975) Bayesian scoring was used to update $\hat{\theta}$, and a Bayesian modal estimate was computed at the end of the test to obtain the final $\hat{\theta}$. Items were adaptively selected from information tables on the basis of maximum information. [An information table consists of lists of items by θ level. Within each list, all the items in the pool (40 in this case) are arranged in descending order of the values of their information functions computed at that θ level. The information tables used in this study were computed for 37 θ levels equally spaced along the (−2.25, +2.25) interval].

ASVAB scores. The Armed Forces Qualification Test (AFQT), a composite score derived from ASVAB subtests, was obtained from the records of most examinees. AFQT scores are used by all the military services to determine eligibility for enlistment. These scores were used to assess the equivalence of the three groups.

Covariance Structure Analysis

The equality of $\hat{\theta}$s calculated from P&P and computer-estimated item parameters was investigated using covariance structure analysis based on the eight variables defined in Table 2. The formal model can be defined as follows. Let a random observation i from Group 3 be denoted as Y_{ti}, where t denotes one of four adaptive subtests (GS, AR, WK, or SI). In the adaptive test, item selection and scoring are assumed to be based on item parameters that are representative of a population of item parameters, where the population consists of parameters obtained from each of a large number of CMOAs. [A large number of hypothetical MOAs can be defined from various combinations of item display format (defined by the choice of font, color, and display medium) and response format (defined by the choice of format of the answer sheet or automated input device).] The random observation is assumed to be on a standardized score scale with a mean of 0.0 and a variance of 1.00. The 1×4 vector of observations, $\mathbb{Y}_i = \{Y_{ti}\}$, is assumed to be from a multivariate normal random variable with a 4×4 correlation matrix, Φ.

A standardized random observation based on the use of item parameters from a specific CMOA is denoted W_{tmi} and is assumed to have a linear regression on Y_{ti},

$$W_{tmi} = \rho_{tm} Y_{ti} + e_{tmi}. \quad (1)$$

The e_{tmi} are errors that are assumed to have a multivariate normal distribution and to be independent of each other and of the Y_{ti}. They are interpreted as errors in test scores due to nonsystematic departure of item parameters from the population-representative item parameters used to obtain Y_{ti}. These errors are a combination of various CMOA effects not definable by a linear transformation of the score scale, such as sampling variation of the parameter estimates and variation due to the interaction of specific item contents and the CMOA. Note that because the W_{tmi} and Y_{ti} are both standardized variables, the regression coefficient, ρ_{tm}, is the correlation between these variables, and the error variance is $1 - \rho_{tm}^2$. Also, note that the equivalence of ρ_{tm} across CMOA for each test can be taken as an indicator of similar amounts of nonsystematic calibration error across CMOA.

From these definitions of W_{tmi} and Y_{ti}, it follows that the observed score on test t in medium m can be written as:

$$X_{tmi} = \sigma_{tm} W_{tmi} + \mu_{tm}, \quad (2)$$

where σ_{tm} and μ_{tm} are the observed scale standard deviation (SD) and location (mean) parameters, respectively. If the CMOA has no linear effect on the score scale for test t, then σ_{tm} and μ_{tm} are the same for all m (i.e., for all CMOA).

The covariance matrix $\boldsymbol{\Sigma}$ among the eight variables can be modeled in terms of several parameter matrices:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\left(\mathbb{R}^{1/2}\mathbb{J}\boldsymbol{\Phi}\mathbb{J}'\mathbb{R}^{1/2} - \mathbb{R} + \mathbb{I}_8\right)\boldsymbol{\Lambda}, \quad (3)$$

where $\boldsymbol{\Lambda}$ and $\mathbb{R}$ are 8×8 diagonal matrices with elements

$$\boldsymbol{\Lambda} = \text{diag}\{\sigma_{GSC}, \sigma_{ARC}, \sigma_{WKC}, \sigma_{SIC}, \sigma_{GSP}, \sigma_{ARP}, \sigma_{WKP}, \sigma_{SIP}\} \quad (4)$$

and

$$\mathbb{R} = \text{diag}\{\rho_{GSC}, \rho_{ARC}, \rho_{WKC}, \rho_{SIC}, \rho_{GSP}, \rho_{ARP}, \rho_{WKP}, \rho_{SIP}\}. \quad (5)$$

The $\boldsymbol{\Lambda}$ matrix contains the SDs of the observed variables and the $\mathbb{R}$ matrix contains the reliability parameters. These reliability parameters measure only one source of error variance: the random error variance in test scores arising from sampling errors in item parameters. These reliability parameters do not measure error in the traditional sense, which measures the error in test scores associated with the sampling of items from an infinite pool of items. The matrix $\mathbb{J}$ is 8×4 with

$$\mathbb{J} = \begin{bmatrix} \mathbb{I}_4 \\ \mathbb{I}_4 \end{bmatrix}, \quad (6)$$

where $\mathbb{I}_4$ is a 4×4 identity matrix, and $\mathbb{I}_8$ denotes an 8×8 identity matrix.

In Equation 3, $\boldsymbol{\Phi}$ is a 4×4 symmetric matrix with diagonal elements equal to 1. The $\boldsymbol{\Phi}$ matrix contains the disattenuated correlations among the four tests. Note that in this context, the correlations are corrected for calibration errors only. These correlations are not corrected for attenuation due to measurement errors.

From Equation 3, the disattenuated correlation matrix among the eight variables is given by:

$$\mathbb{J}\boldsymbol{\Phi}\mathbb{J}' = \begin{bmatrix} \boldsymbol{\Phi}_{CC} & \boldsymbol{\Phi}_{PC} \\ \boldsymbol{\Phi}_{CP} & \boldsymbol{\Phi}_{PP} \end{bmatrix}, \quad (7)$$

where the three nonredundant submatrices are constrained by the model to be equivalent: $\boldsymbol{\Phi}_{CC} = \boldsymbol{\Phi}_{PC} = \boldsymbol{\Phi}_{PP} (= \boldsymbol{\Phi})$. From classical test theory, the product $\mathbb{R}^{1/2}\mathbb{J}\boldsymbol{\Phi}\mathbb{J}'\mathbb{R}^{1/2}$ represents the correlation matrix among observed variables, with the eight reliability parameters along the diagonal. Consequently, the sum $\mathbb{R}^{1/2}\mathbb{J}\boldsymbol{\Phi}\mathbb{J}'\mathbb{R}^{1/2} - \mathbb{R} + \mathbb{I}_8$ represents the correlation matrix among observed variables, with 1s in the diagonal. Finally, by pre- and post-multiplying the observed correlation matrix by $\boldsymbol{\Lambda}$ (the 8×8 diagonal matrix of SDs), the observed covariance matrix $\boldsymbol{\Sigma}$ is obtained.

In addition to estimating the model given by Equation 3, an additional model was examined to test the equivalence of the reliability parameters across the CMOA. The constraints imposed by the two models are summarized in Table 3. Model 1 imposed Constraint A, which equated the disattenuated correlations across the CMOA; Model 2 imposed both Constraints A and B, where B constrained the reliability parameters (see Table 3). Consequently in Model 2, the reliability values for each test were constrained to be equivalent across the two calibration media. Model parameters were estimated by normal theory maximum likelihood using the SAS procedure CALIS (SAS Institute Inc., 1990).

Table 3
Model Constraints: Model 1 Imposed Constraint A, and Model 2 Imposed Constraints A and B

Constraint	Parameters
A	$\boldsymbol{\Phi}_{CC} = \boldsymbol{\Phi}_{PC} = \boldsymbol{\Phi}_{PP}$
B	$\rho_{GSC} = \rho_{GSP}, \rho_{ARC} = \rho_{ARP}, \rho_{WKP} = \rho_{WKC}, \rho_{SIC} = \rho_{SIP}$

Models 1 and 2 represent a hierarchy of nested models. Consequently, the χ^2 difference test can be used to examine the statistical significance of each set of constraints. Significance tests were performed on each set of constraints listed in Table 3. For both models, the likelihood ratio χ^2 statistic of overall fit was calculated. To test the equivalence of disattenuated correlations across the CMOA ($\boldsymbol{\Phi}_{CC} = \boldsymbol{\Phi}_{PC} = \boldsymbol{\Phi}_{PP}$), the likelihood χ^2 value for Model 1 was used. To test the equivalence of the reliability parameters, the difference between the χ^2 values of Models 1 and 2 was evaluated. Under the null hypothesis, this difference

was distributed as χ^2 with 4 degrees of freedom (df).

Results

Group Equivalence

Two cases in Group 3 had fewer than 10 valid responses for WK and SI and consequently were eliminated from all subsequent analyses for these two tests. Thus, the sample sizes for Group 3 were 988 for GS and AR and 986 for WK and SI. An ANOVA indicated a nonsignificant difference among the three group means on AFQT. This result provided some assurance that the three groups were randomly equivalent with respect to ASVAB aptitudes.

Difficulty Parameter Comparison

A comparison of the IRT difficulty parameters across the two media for Groups 1 and 2 provided one assessment of the effects of using alternative CMOA on the score scale. Ideally, the parameters from the two media should fall along a diagonal (45°) line. Systematic effects on the score scale would cause the points to fall along a different line (if linearly related), or curve (if nonlinearly related). Nonsystematic effects would influence the degree of scatter about the line.

Figures 1a–1d display the plot of difficulty parameters estimated from the two CMOA, for each of the four tests. As each plot indicates, the parameters fell along the diagonal with a small degree of scatter. This result is consistent with small or negligible effects of the calibration media on the score scale.

Covariance Structure Analysis Results

The sample correlation matrix among the eight $\hat{\theta}$s for Group 3 is displayed in Table 4. Also displayed in Table 4 are the means and SDs of these variables.

The estimated parameters of Model 1 are displayed in Tables 5 and 6. As indicated by the $\hat{\rho}$ columns of Table 6, the reliability values for both CMOA

Figure 1
Difficulty Parameter Estimates from the Two CMOA

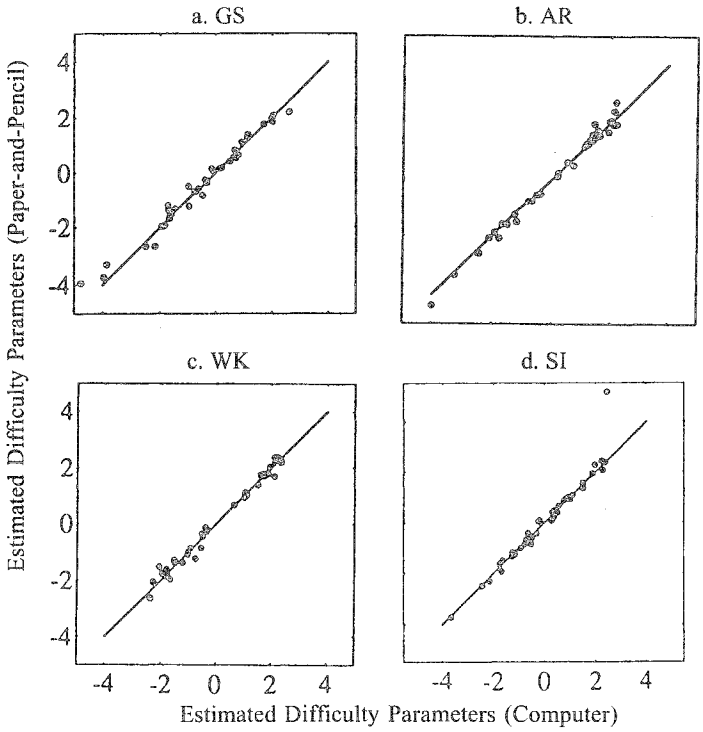

Table 4
Correlations, Means, and Standard Deviations (SDs) Among the Eight Scores for Group 3

Variable and Statistic	Variable							
	X_{GSC}	X_{ARC}	X_{WKC}	X_{SIC}	X_{GSP}	X_{ARP}	X_{WKP}	X_{SIP}
X_{GSC}								
X_{ARC}	.504							
X_{WKC}	.734	.446						
X_{SIC}	.601	.354	.496					
X_{GSP}	.970	.506	.728	.587				
X_{ARP}	.507	.981	.449	.351	.506			
X_{WKP}	.737	.450	.980	.500	.730	.451		
X_{SIP}	.605	.351	.490	.956	.587	.349	.494	
Mean	.025	−.027	.012	.042	.069	−.068	.034	.012
SD	.857	.927	.877	.866	.863	.947	.853	.896

were nearly 1.0. These results indicate that a very small amount of random error among test scores was attributable to estimation errors among item parameters. The estimated σ values for each CMOA are provided in the last two columns of Table 6.

Table 5
Estimated Disattenuated Correlation Matrix $\hat{\Phi}$ for Model 1

Test	Test			
	GS	AR	WK	SI
GS	1.00			
AR	.52	1.00		
WK	.75	.46	1.00	
SI	.62	.36	.51	1.00

The results of overall fit for Models 1 and 2 are displayed in Table 7. As indicated in Table 7, the likelihood ratio χ^2 value for Model 1 was nonsignificant, which provides support for the equivalence of the disattenuated correlation matrices: $\Phi_{CC} = \Phi_{PC} = \Phi_{PP}$. This result indicates that CMOA did not alter the constructs measured by the four tests.

The χ^2 test based on the differences between Models 1 and 2 indicated no difference between the reliability parameters across the two calibration media ($\chi^2 = 19.267 - 14.066 = 5.201$, $df = 18 - 14 = 4$, $p = .27$). This result supports the contention that the reliability of CATs is independent of the medium used to calibrate the item parameters.

Table 6
Estimated Reliabilities $\hat{\rho}$ and Standard Deviations $\hat{\sigma}$ for Model 1

Test	$\hat{\rho}$		$\hat{\sigma}$	
	Computer	P&P	Computer	P&P
GS	.983	.958	.857	.863
AR	.978	.985	.927	.947
WK	.976	.984	.877	.853
SI	.956	.957	.866	.896

Table 7
χ^2 Values for Tests of Overall Fit of Models 1 and 2

Model	Constraints	*df*	χ^2	*p*
1	A	14	14.066	.44
2	A, B	18	19.267	.38

Conclusions

The good fit of Model 1 to the data indicated that, for the four tests, the disattenuated correlations among the scores based on the computer-based calibration, Φ_{CC}, did not differ significantly from the disattenuated correlations among the scores based on the P&P-based calibration, Φ_{PP}; and neither of these sets of correlations differed significantly from the disattenuated cross-correlations of scores based on the two types of calibration, Φ_{PC}. This is consistent with the lack of within-trait medium-of-administration correlational effects found by Mead & Drasgow (1993). It also extends the conclusions drawn by Mead and Drasgow to the consistency of disattenuated correlations between traits.

The results from the comparison of Models 1 and 2 indicated that, for the four tests, equal amounts of nonsystematic error variance ($1 - \rho^2_{tm}$) were obtained with the use of the computer-based and P&P-based item calibrations. This is generally consistent with—and extends—the findings of Divgi (1986) and Divgi & Stoloff (1986), in which the

computer-based calibration was based primarily on data from adaptively administered items.

The secondary effect under investigation was the influence of calibration medium on the score scale. A comparison of the difficulty parameters across the two media indicated very little or no distortion in the scale. For all four tests, the difficulty parameters tended to fall along a diagonal (45°) line.

An important practical implication of the results of this study is that item parameters calibrated from a P&P administration of items can be used in power CATs of cognitive constructs—such as those found on the CAT-ASVAB—without changing the construct being assessed and without incurring lower reliability. Although the descriptive analyses of difficulty parameters suggest little or no effect of calibration medium on the score scale, Green et al. (1984) noted that if scale effects do exist, they can be corrected by equating to a reference form that defines the score scale to be used for selection and classification decisions. When this is done, distortions in the mean, variance, and higher moments of the observed scores have no effect on selection and classification decisions.

References

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

Divgi, D. R. (1986). *Determining the sensitivity of CAT-ASVAB scores to changes in item response curves with the medium of administration* (Report No. 86-189). Alexandria VA: Center for Naval Analyses.

Divgi, D. R., & Stoloff, P. H. (1986). *Effect of the medium of administration on ASVAB item response curves* (Report No. 86-24). Alexandria VA: Center for Naval Analyses.

Green, B. F., Bock, R. D., Linn, R. L., Lord, F. M., & Reckase, M. D. (1984). A plan for scaling the computerized adaptive Armed Services Vocational Aptitude Battery. *Journal of Educational Measurement, 21,* 347–360.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement.* Homewood IL: Dow Jones-Irwin.

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

Moreno, K. E., & Segall, D. O. (1992). CAT-ASVAB precision. *Proceedings of the 34th Annual Conference of the Military Testing Association, 1,* 22–26.

Moreno, K. E., Segall, D. O., & Kieckhaefer, W. F. (1985). A validity study of the computerized adaptive testing version of the Armed Services Vocational Aptitude Battery. *Proceedings of the 27th Annual Conference of the Military Testing Association, 1,* 29–33.

Moreno, K. E., Wetzel, D. C., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests. *Applied Psychological Measurement, 8,* 155–163.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive testing. *Journal of the American Statistical Association, 70,* 351–356.

Prestwood, J. S., Vale, C. D., Massey, R. H., & Welsh, J. R. (1985). *Armed Services Vocational Aptitude Battery: Development of an adaptive item pool* (AFHLR-TR-85-19; Technical Rep. No. 85-19). Brooks Air Force Base TX: Air Force Human Resources Laboratory.

SAS Institute Inc. (1990). *SAS/STAT user's guide* (4th ed.). Houston TX: Author.

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement, 26,* 261–271.

Weiss, D. J. (Ed.). (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide.* Princeton NJ: Educational Testing Service.

Acknowledgments

The opinions expressed in this article are those of the authors, are not official, and do not reflect the views of the Department of the Navy or the Department of Defense. The Department of Defense is granted a nonexclusive right to reproduce this article without royalty for government purposes.

Author's Address

Send requests for reprints or further information to Rebecca D. Hetter, Personnel Systems Department (Code 12), Navy Personnel Research and Development Center, 53335 Ryne Road, San Diego CA 92152-7250, U.S.A.