THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, *4*(4), 297–313

FORMAL PAPERS

Predicting Aircraft Pilot-Training Success: A Meta-Analysis of Published Research

David R. Hunter
Federal Aviation Administration, Washington, DC

Eugene F. Burke
London Fire and Civil Defence Authority

Results are given from a meta-analysis of validities for aircraft pilot-selection measures. Sixty-eight published studies were identified for the 1940-to-1990 period, from which 468 correlations were extracted for a cumulated sample of 437,258 cases. The method proposed by Hunter and Schmidt (1990b) was applied to produce a bare-bones analysis. Mean sample-weighted correlations, estimates of true variance, and confidence intervals were computed. Several classes of predictors were found to have confidence intervals that did not include zero, indicating possible generalizability of validities. For the most part, however, the variance accounted for by sampling error alone was small. The effects of moderator variables (including nationality, service, decade of publication, and aircraft type) were evaluated. Of these, decade of publication was most consistently correlated with obtained validities and was associated with a decline in average validities over the five decades of studies examined. Limitations on interpretation of the results and problems associated with the analysis and interpretation of data from the published reports are discussed, and the range of correlations that might be expected from a composite of the groups of predictors that were examined is reported.

Requests for reprints should be sent to David R. Hunter, Office of Aviation Medicine, AAM–240, Federal Aviation Administration, 800 Independence Avenue SW, Washington, DC 20591.

Aviator selection historically has been an area of great interest and considerable research effort. For various reasons, the bulk of the published reports has dealt with selection for *ab initio,* military pilot training, although much of what has been accomplished in the military setting is directly applicable to civil aviation. For the most part, military training utilizes applicants with little or no previous experience in aviation, and trains them to a criterion performance level in about 200 flying hr over a 12-month period. Using these flight-naive trainees and a stringent timetable results in a significant number of training failures. Typical training attrition rates over the last 20 years have been on the order of 25%, with an average cost for each failure ranging from $50,000 to $80,000 for the U.S. Air Force (Hunter, 1989; Siem, Carretta, & Mercatante, 1987). Because of these costs, the military services of the United States and other countries have conducted continuous, wide-ranging studies to identify and evaluate new selection measures. However, results from individual studies often seem contradictory in the range and signs of validities (the correlations between predictor and criterion scores) for individual selection measures. A narrative review of those efforts is provided in Hunter (1989).

The 1980s saw a growth in publications of quantitative research reviews with the development of meta-analytic methods. In the context of personnel selection, these methods have come to be known as *validity generalization* (VG), and this article uses the approach of VG's principal proponents, Hunter and Schmidt (1990b). The advantage of VG is the explicit evaluation of factors that may influence the size of a validity estimate in a specific study. By cumulating data from several studies on both validity and study characteristics, it is possible to estimate the generalizability of a selection measure across different selection contexts. Knowledge of the reliabilities of predictor and criterion measures, the degree of selection present in a validation sample (i.e., degree of range restriction), and the sample size of the study allows psychometric analysis of the extent to which apparent variation across studies is due merely to the limitations of individual study design (i.e., due to the effects of artifacts such as variation in sample sizes or predictor and criterion reliabilities). The basic steps in a VG analysis are first to compute a weighted mean validity across studies and the variance about that mean. Total variance may then be decomposed into components attributable to artifacts and to true variation. When the mean validity is nonzero and true variance about that mean is zero, then the selection construct or method of interest is said to be generalizable.

The analysis reported in this article is based on a review of articles published in the literature and in the technical reports of defense organizations. The extent to which the sample of studies so identified is representative of the full population of civilian and military selection studies is unknown. Some bias may well be present, given that the review focused only on articles and reports in English, and that those identified are predominately North American and British. The analysis was also confronted by limited information on study artifacts, in that the majority of studies did not report details of predictor or criterion reliabilities or the degree of range restriction. Although distributions of artifacts may be estimated from what information is available, this paper presents a bare-bones VG analysis in which the only artifact included is variation due to study sample size. A further complication was that military studies predominately used pass–fail in training as the criterion measure. Dichotomization (or discontinuation) of a variable places a limit on the maximum correlation that can be expected. For example, with a 50–50 split (giving the maximum variance in a dichotomized variable) the maximum correlation is 0.798 and not 1.0 (as the split moves away from 50–50, so the maximum correlation decreases; see Cohen, 1983; Hunter & Schmidt, 1990a). Despite these problems, the cumulative sample sizes identified through the literature review are substantial and allow for analyses that more accurately clarify and represent the underlying relations between predictor measures and training performance criteria than is possible in any of the individual studies.

METHOD

Sample

Studies were identified through manual and computerized searches of the *Psychological Abstracts* for the years 1940 through 1990 and through manual searches of the published bibliographies of the military services. Promising sources cited in studies obtained in this manner were also reviewed. The studies so identified included those from both refereed and nonrefereed sources (primarily military reports) and unpublished studies. Studies were carefully examined to ensure that there were no duplications, particularly between the military and refereed sources. If the same study was reported in more than one source (e.g., as both a military technical report and a journal article), the more generally available citation was used.

Studies were selected for inclusion if they reported a correlation between one or more predictors and one or more flying-training criteria. There were 68 studies so identified for the 1940-to-1990 period. The studies included in the analyses are listed in the Appendix; however, space limitations preclude giving full details of the studies and the measures. That information is available as a technical report (Hunter & Burke, 1990), and the complete data base from which this study was conducted is available in a computer-readable format from the first author.

Procedure

Each study was reviewed and the correlation, sample size, and other information regarding potential artifacts for each study was coded and recorded in a data base. In many cases, several correlations were reported in a study utilizing a single group of subjects. In those instances in which the correlations were related to a single predictor construct or instrument (e.g., multiple performance measures taken from a flight simulator or several scores derived from performance on a single test) a composite was created by converting the correlations to Fisher's *Z* and then averaging. This single, averaged correlation was then entered into the VG analysis to represent the contribution of that construct or measure. This process reduced the sample of correlations from 664 to 501. In those cases in which the same examinees supplied responses to different tests measuring different constructs in a single study (e.g., in a multiple-test battery) each correlation was entered separately. We recognized the nonindependence of the data that this treatment produces, but we felt that this approach was more attractive than ignoring the available data.

This sample was further reduced by eliminating from the analyses any correlations reported for combinations of predictors (e.g., multiple-test batteries) in which only the correlation for the composite score was given. This final screening reduced the sample to 468 validities, which were entered into the analysis reported herein. The distribution of study characteristics for these correlations is given in Table 1.

Finally, the signs of error-scored measures (e.g., psychomotor coordination) were reflected, so that a positive correlation would indicate that superior test performance is associated with superior performance in training. An exception to this treatment, however, was that given to personality measures: Because there was no a priori expectation regarding the direction of prediction of these measures (e.g., whether one should expect superior flying performance to be associated with high or low authoritarianism) their signs were not changed. Additional research may address alternative treatments of this problem (cf. Tett, Jackson, & Rothstein, 1991).

Data Analysis

For reasons presented earlier, the VG analysis reported in this article only incorporated the artifact of sampling error due to variation in sample size cross-studies. This is referred to as a *bare-bones analysis,* although Hunter and Schmidt (1990b) stated that sampling error was the most significant of the 11 artifacts that they listed.

Four parameters were computed for each predictor class using formulae from Hunter and Schmidt (1990b): (a) the sample-weighted mean validity, (b) the sample-weighted variance in validities about that mean (observed variance), (c) variance in validities due to sampling error (error variance), and (d) corrected variance (true variance). Subtracting error variance from observed variance provides the true variance estimate. The equations

TABLE 1
Distribution of Study Characteristics

Study Characteristic	*Number of Correlations*	*N*
Sample service		
Air Force	289	351,317
Navy	110	62,098
Army	42	20,218
Civilian	27	3,625
Sample nationality		
United States	362	408,724
United Kingdom	20	3,108
Canada	52	9,743
Other	34	15,683
Aircraft type		
Fixed wing	402	413,176
Rotary wing	66	22,082
Criterion category		
Dichotomous (pass–fail)	391	404,969
Continuous	77	32,289
Total	468	437,258

(Hunter & Schmidt, 1990b) required to compute these parameters were coded in the programming language of the data base management system (Microsoft Access®; Microsoft Corporation, Redmond, WA), which was used to maintain the data base. This system was then used to select cases for analysis and to make all VG calculations.

VG decision rules. The estimate of true variance can be used to apply two decision rules in classifying the validity generalization of a predictor group, in which a *predictor group* represents measures of a construct such as verbal or quantitative ability. The first decision rule, the 75% rule, is taken from Hunter and Schmidt's (1990b) experience in VG analyses; they have found that sampling error tends to account for 75% of the variance in validities when in fact the true variance is zero. Thus, in a bare-bones analysis, if the ratio of error to observed variance is 75% or greater, then observed variance is said to be entirely attributable to artifacts. That is, once artifacts or the limitations within particular studies are taken into account, then there is no true variation and validity is generalizable. When this ratio is less than 75%, then sufficient true variation remains to warrant further analyses to identify the sources of that variation (i.e., significant moderators).

The second decision rule is to apply a confidence interval about the mean validity across individual studies. In fact, two such intervals can be computed: a traditional confidence interval using the square root of the observed variance or a credibility limit computed using the square root of the true variance. The latter limit is based on the distance between the means of the probability distributions for null and alternate hypotheses and, for reasons outlined by Hunter and Schmidt (1990b) and Whitener (1990), a lower, 90% limit is used to determine whether the mean validity falls within the null distribution with mean validity of zero. That is, if the 90% credibility limit is greater than zero, then the mean validity across studies can be assumed to be nonzero.

Both rules were combined in the analysis to give four classes of results:

1. 75% rule positive and 90% confidence limit positive: No true variation in studies. Validity not significantly influenced by moderator variables. Mean validity generalizable.
2. 75% rule negative and 90% confidence limit positive: Validity nonzero, but size of validity influenced by moderators.
3. 75% rule negative and 90% confidence limit negative: Uncertainty about true validity of predictor. Validity may be nonzero in a subset of studies or contexts. Validity not generalizable.
4. 75% rule positive and 90% confidence limit negative: True validity is zero. This would be immediately apparent from a mean validity of zero and little variation about that mean.

Analysis of moderators. The procedure used for analyzing the impact of moderator variables is akin to that recently reported by Mead and Drasgow (1993). First, moderators identifiable from the reports reviewed were coded as dummy variables. For example, aircraft type was coded 1 for fixed wing and 0 for rotary wing, service was coded 1 for military and 0 for civilian, nationality was coded 1 for U.S. and 0 for other, and decade of study was coded 1 for the 1940-to-1960 period and 0 for the 1961-to-1990 period. Thus, each study was assigned a 1 or 0 code for each of these four potential moderator variables. The following model testing procedure (taken from Dwyer, 1983) was then applied separately for those predictors falling into Class 2 under the VG decision rules just described.

In Step 1, a saturated regression model including all moderators was computed. That is, validities for a predictor group (e.g., spatial ability) were regressed onto all four moderator variables. In Step 2, four restricted models were computed with each moderator removed in turn. The difference between the R^2 values for the saturated and restricted models then gave a direct estimate of the impact of a moderator on the validity observed for a given group of predictors. The difference in R^2 can also be tested for statistical significance using an F ratio (for details, see Dwyer, 1983).

These analyses were carried out using Statistical Package for the Social Sciences (SPSS; SPSS, Inc., Chicago) for Windows (Version 6.0).

RESULTS

Table 2 summarizes the VG results for the predictor groups. What is immediately apparent is that only a small percentage of observed variation in validities is accounted for by sampling error alone, the highest percentage being the 45% found for measures of fine dexterity. As such, none of the predictor groups satisfied the 75% rule, placing them all into either Class 2 or Class 3 under the decision rules described earlier. The 90% credibility limits on the far right of Table 2 serve to distinguish between predictors as being moderated (Class 2) or nongeneralizable (Class 3). Those falling into Class 3 are verbal ability, fine dexterity, education attainment, and personality.

Moderator analyses were pursued for the predictors falling into Class 2 within the scope possible for such analyses. Both age and reaction time were excluded from further analyses due to the small number of validities available for those predictor groups. Table 3 summarizes the moderator analyses for the other predictor groups in Class 2.

With respect to the four moderators included in the analyses, decade of study appears to have the most significant impact. Data for five of the predictor groups are further decomposed in Table 4, from which it may be seen that there has been a general decline in validity since 1961. The earlier studies, dominated by research during World War II, are also notable for

TABLE 2
VG Results by Predictor Group

Predictor	r_{mean}	N_x	N_s	$\hat{\sigma}_r^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_p^2$	$\hat{\sigma}_{explained}^2$	L_{95}	U_{95}
General ability	0.13	14	8,071	.008	.002	.006	21	−0.05	0.30
Verbal ability	0.12	17	22,841	.012	.001	.011	6	−0.09	0.33
Quantitative ability	0.11	34	46,884	.003	.001	.002	28	0.01	0.21
Spatial ability	0.19	37	52,153	.005	.001	.004	14	0.05	0.32
Mechanical	0.29	36	42,418	.009	.001	.008	8	0.11	0.48
General information	0.25	13	29,951	.010	.000	.009	4	0.06	0.44
Aviation information	0.22	23	25,295	.007	.001	.006	12	0.06	0.38
Gross dexterity	0.32	60	48,988	.007	.001	.006	13	0.15	0.49
Fine dexterity	0.10	12	2,792	.009	.004	.005	45	−0.09	0.29
Perceptual speed	0.20	41	33,511	.006	.001	.005	19	0.05	0.35
Reaction time	0.28	7	10,633	.004	.001	.003	16	0.16	0.39
Biodata inventory	0.27	21	27,004	.011	.001	.010	6	0.07	0.47
Age	−0.10	9	13,810	.006	.001	.005	11	−0.25	0.05
Education	0.06	9	6,163	.012	.001	.011	12	−0.16	0.27
Job sample	0.34	16	2,814	.012	.005	.007	37	0.19	0.55
Personality	0.10	46	22,486	.018	.002	.016	11	−0.16	0.37

Notes. r_{mean} = mean sample-weighted correlation, N_x = number of validities, N_s = total sample size across studies, $\hat{\sigma}_r^2$ = observed variance, $\hat{\sigma}_e^2$ = error variance, $\hat{\sigma}_p^2$ = corrected variance, $\hat{\sigma}_{\text{explained}}^2$ = percentage of variance explained, L_{95} = lower 95% confidence interval, U_{95} = upper 95% confidence interval.

TABLE 3
Moderator Analyses

Predictor	R^2 *Saturated Model*	*Decade Removed*	*Service Removed*	*Nationality Removed*	*Aircraft Type Removed*
Quantitative ability	0.109	0.106	0.108	0.049	0.042
Spatial ability	0.246	0.109[a]	0.189	0.246	0.245
Mechanical	0.513	0.070[b]	0.508	0.499	0.513
Aviation information	0.361	0.285	0.284	0.291	0.361
Gross dexterity	0.474	0.220[b]	0.277[b]	0.473	0.214[b]
Perceptual speed	0.299	0.053[b]	–	0.282	0.248
Biodata inventory	0.583	0.056[b]	0.544	0.552	–
Job sample	0.576	0.533	0.565	0.010[b]	0.576

[a] Difference between saturated and restricted model significant at the .05 level. [b] Difference significant at the .01 level.

TABLE 4
Analysis by Decade for Five Predictors

Predictor	r_{mean}	N_x	N_s	δ_r^2	δ_e^2	$\delta^2_{explained}$	C_{90}
1940–1960							
Spatial ability	0.21	14	41,693	.005	.080	10	0.12
Mechanical	0.32	20	35,619	.004	.003	12	0.23
Gross dexterity	0.35	24	37,177	.004	.004	14	0.25
Perceptual speed	0.23	7	25,343	.001	.001	17	0.18
Biodata inventory	0.30	12	22,575	.004	.004	10	0.20
1961–1990							
Spatial ability	0.12	23	10,460	.006	.004	38	0.02
Mechanical	0.14	16	6,799	.008	.006	27	0.01
Gross dexterity	0.22	36	11,811	.005	.002	60	0.15
Perceptual speed	0.11	34	8,168	.008	.004	49	0.01
Biodata inventory	0.09	9	4,429	.005	.003	41	0.00

TABLE 5
Analysis by Service, Aircraft Type, and Nationality for Two Predictors

Predictor	r_{mean}	N_x	N_s	δ_r^2	δ_e^2	$\delta^2_{explained}$	C_{90}
Moderator code = 1							
Gross dexterity[a]	0.33	47	44,620	.007	.006	12	0.20
Gross dexterity[b]	0.33	52	45,126	.008	.007	12	0.19
Job sample	0.29	10	1,846	.002	.000	100	0.29
Moderator code = 0							
Gross dexterity	0.26	13	4,368	.005	.002	55	0.19
Gross dexterity	0.27	8	3,862	.002	.001	76	0.22
Job sample[c]	0.44	6	968	.015	.004	27	0.27

[a] Moderator = service (1 = Air Force, 0 = Other). [b] Moderator = aircraft type (1 = fixed wing, 0 = rotary wing). [c] Moderator = nationality (1 = United States, 0 = other).

much larger average sample sizes. Consequently, in addition to lower validities, research in more recent decades is also characterized by larger sampling error, as indicated by the fact that sampling error accounts for more of the observed variance in validities. As such, estimates of validity are more variable due to these smaller study samples. Of the predictors, biodata inventories show the greatest decline (from 0.30 to 0.09) and is the only group to show a 90% credibility limit of zero for later decades.

The situation regarding gross dexterity is complicated by the fact that three moderators were found to have significant relations to validity. Decomposing the data by service shows a lower mean validity and greater sampling error for non–Air Force studies and for studies involving rotary-wing aircraft. Decomposing job sample measures by nationality shows a lower mean validity and lower variability for U.S. studies. However, both predictors show 90% credibility values substantially above zero despite moderator breakdowns, thus confirming their generalizability. Both sets of analyses are summarized in Table 5.

Finally, we offer a comment on the results for quantitative ability and aviation information measures. The low multiple correlations obtained from the saturated models merely shows that this study was unsuccessful in identifying those moderators that influence the validity of these predictors. This was due, as we note in the Discussion section, to the limited information on study characteristics given in published reports. At present, although the validity of these predictors is moderated, variables explaining this moderation have yet to be identified.

DISCUSSION

Table 6 summarizes the results obtained from the VG analysis of pilot validities. One conclusion is clear: Although none of the predictor groups fell into Class 1, neither did any fall into Class 4 (no validity). The pattern of results, however, is complex, with the validity of predictors falling into Class 2 influenced by moderator variables. This study has considered four such variables, and the most significant was found to be decade of study. Why should time influence validity; or rather, why should validity decline with time? One may suggest that such a decline is attributable to changes in the population of applicants (education, experience, and more selective recruitment contributing to increased range restriction in estimates) or changes in the training/operational environment (i.e., the criterion predicted). More information than that generally found in the reports reviewed is required before such hypotheses can be evaluated.

Indeed, as seems to have become a common plea in meta-analysis reviews, a general improvement in the quality of study information is needed to permit a more exhaustive review of pilot validities. Among the data required are information on the reliability of measures; the method by which

TABLE 6
Classification of Results

Class and Description	*Predictor Measures*
1: Validity generalizable	None
2: Validity moderated	Quantitative ability, spatial ability, mechanical, aviation information, general information, gross dexterity, perceptual speed, reaction time, biodata inventory, and job sample
3: Validity not generalizable	General ability, verbal ability, fine dexterity, age, education, and personality
4: No validity	None

reliabilities were estimated; and, where a dichotomized criterion was used, the nature of any nominal classifications and percentages falling into categories. Also important is better reporting of criterion data over time and raters to gauge the impact of criterion quality on validity. It seems that, as ever, although effort is often put into the predictor side of the equation, the criterion remains a Cinderella factor despite the distortion that inadequate criteria have on estimates of the cost benefit of predictor investment.

The mean validities reported here, even for those sets of measures falling into Class 2, may appear small. However, because they are subject to range restriction and dichotomization (for which they are uncorrected), they should be interpreted with some caution, because they are likely to underestimate true validity. Thorndike's (1949) description of research for the U.S. Army Air Force during World War II provides comparisons of validities prior to and after the impact of range restriction. A selection test battery was developed for pilots, and data were collected under conditions in which all applicants were selected regardless of their test scores. The calculation of validities was therefore straightforward, as there were minimal effects of range restriction. The results showed the validities for seven tests to range from 0.18 (finger dexterity) to 0.46 (general information), with a composite validity of 0.64. A retrospective analysis was then performed in which only those with composite scores exceeding the subsequent cutoff set for use of the test battery were included (only 13% of the original sample met the cutoff). The composite validity then fell to 0.18. Of interest is the effect of range restriction on the validity of an obvious predictor, complex coordination (a psychomotor measure). In the original unselected sample, the validity was found to be a respectable 0.40. In the selected sample, it fell to –0.03, emphasizing the need to consider artifacts in evaluating the utility of predictors in pilot selection.

Such an effect raises the issue of sample sizes required for adequate estimation of validities. Some typical values were taken to perform a power analysis for this purpose, the results of which are given in Table 7. A hiring ratio of 30% (i.e., only the top 30% of applicants are selected), and a pass–fail ratio of 70:30 were used. (These values were suggested by the second author's experience.) Three correlation values were taken from

Cohen (1977) to represent large, medium, and small effect sizes: 0.5, 0.3, and 0.1, respectively. As can be seen from Table 7, taken individually, the artifacts considerably reduce the observed correlation to be expected. Taken in combination, the effect on the expected correlation is dramatic. The minimum sample sizes given are for a one-tailed Type I error rate of 0.05 and a Type II error rate of 0.2 (i.e., 80% statistical power) and the expected validity value given under the combined condition (i.e., subject to both range restriction and dichotomization). The acquisition of such sample sizes will present difficulties for many individual studies or require a considerable time to amass. Analyses such as that reported herein may therefore be more feasible as a guide to inclusion of predictors in a selection battery. But, as we noted earlier, the value of future VGs of predictors of pilot success will depend on the quantity and quality of information provided by individual studies on predictors, criteria, sample characteristics, and artifacts.

As was described in the Method section, one note of caution needs to be raised in conjunction with interpretation of the result for personality measures. With the recent advent of the Big Five personality taxonomy (Digman, 1989, 1990), there have been some encouraging validities for personality dimensions. Of particular importance in evaluating the validity of a personality scale is the direction of intended prediction. With ability test scores positively scaled, the direction of the hypothesis test is obvious (higher scores are associated with higher performance). Such is not necessarily the case with personality scales. Another concern is with the artifact of factorial (or construct) validity. That is, use of a taxonomy such as the Big Five requires evidence to support the placing of a scale into a particular class of personality dimension. Several studies reported measures that are difficult to

TABLE 7
Sample Size Requirements by Effect Size and Artifact

Original Effect Size	*Individual Range Restriction*	*Artifact Dichotomization*	*Artifacts Combined*	*Sample Required*[a]
0.5 (large)	0.25	0.38	0.19	170
0.3 (medium)	0.14	0.22	0.11	510
0.1 (small)	0.04	0.08	0.03	6,870

[a] Rounded up to the nearest 10. The value gives the sample required to estimate validity under the combined condition.

TABLE 8
Expected Values for a Battery of Class 2 Predictors

Intercorrelation	L_{95}	M	U_{95}
0.30	0.15	0.41	0.67
0.40	0.13	0.36	0.60
0.50	0.12	0.33	0.55

classify either through lack of descriptive information or of explicit reference to a theoretical basis for the measures. Information provided by individual studies as to the direction of prediction and the construct validity of personality measures should be provided in future studies. This would then allow methods such as that proposed by Tett et al. (1991) to be applied in evaluating predictions from such instruments. However, as noted by Burke (1993) with reference to cultural differences in the semantics of personality dimensions, there is evidence to support the placement of personality predictors into Class 3. Martinussen and Torjussen (1993) reported a small-scale evaluation of the Defense Mechanism Test (DMT), a projective instrument; they showed that, although small validities were obtained from Scandinavian studies, studies in the United Kingdom and the Netherlands obtained validities of zero. It may well be, then, that even with better construct data and improved analytical techniques, differences in cultural context (e.g., English speaking versus non–English speaking) may act to moderate validities for specific personality instruments.

Although the mean validities reported in this article may well underestimate the true validities of several predictors, those falling into Class 2 do indicate potential for substantial cost benefits in pilot selection. To estimate these benefits in the form of a battery made up of the predictors falling into Class 2, estimates of composite validity were computed using three levels of average predictor intercorrelation: 0.3, 0.4, and 0.5. Using these intercorrelations and the formula for a composite given by Guilford (1954), three levels of composite validity were calculated: a lower 95% expected value, a mean value, and an upper 95% expected value. The results of these calculations are shown in Table 8. Again, note that these estimates are subject to range restriction. That is, they may underestimate the true validity of a battery of Class 2 predictors. Future VG analyses with access to the intercorrelations among predictor groups would clearly aid in clarifying the best mix that maximizes the benefits suggested by our analyses.

ACKNOWLEDGEMENTS

Eugene Burke was previously Principal Psychologist, Royal Air Force Recruitment and Selection, Science 3 (Air), U.K. Ministry of Defence.

The opinions expressed herein are those of the authors and do not represent the official policies or views of the U.S. Government or the U.K. Ministry of Defence.

REFERENCES

Burke, E. F. (1993). Pilot selection in NATO: An overview. In R. S. Jensen & D. Neumeister (Eds.), *Proceedings of the Seventh International Symposium on Aviation Psychology* (pp.

373–378). Columbus: Ohio State University.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York: Academic.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249–253.

Digman, J. M. (1989). Five robust trait dimensions: Development, stability and utility. *Journal of Personality, 57,* 195–214.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41,* 417–440.

Dwyer, J. H. (1983). *Statistical models for the social and behavioral sciences.* New York: Oxford University Press.

Guilford, J. P. (1954). *Psychometric methods.* New York: McGraw-Hill.

Hunter, D. R. (1989). Aviator selection. In M. F. Wiskoff & G. Rampton (Eds.), *Military personnel measurement* (pp. 129–167). New York: Praeger.

Hunter, D. R., & Burke, E. F. (1990). *An annotated bibliography of the aircrew selection literature* (Research Rep. No. 1575). Alexandria, VA: U.S. Army Research Institute.

Hunter, J. E., & Schmidt, F. L. (1990a). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology, 75,* 334–349.

Hunter, J. E., & Schmidt, F. L. (1990b). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Martinussen, M., & Torjussen, T. (1993). Does DMT (Defense Mechanism Test) only predict pilot performance in Scandinavia? In R. S. Jensen & D. Neumeister (Eds.), *Proceedings of the Seventh International Symposium on Aviation Psychology* (pp. 398–403). Columbus: Ohio State University.

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

Schmidt, F. L., Hunter, J. E., & Urry, V. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61,* 478–485.

Siem, F. M., Carretta, T. R., & Mercatante, T. A. (1987). *Personality, attitudes, and pilot training performance: Preliminary analysis* (Tech. Rep. No. AFHRL-TR-87-62). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44,* 703–742.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques.* New York: Wiley.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility limits in meta-analysis. *Journal of Applied Psychology, 75,* 315–321.

APPENDIX: STUDIES USED IN THE META-ANALYSIS

Ambler, R. K., Bair, J. T., & Wherry, R. J. (1960). Factorial structure and validity of naval aviator selector variables. *Aerospace Medicine, 31,* 456–461.

Ambler, R. K., Johnson, C. W., & Clark, B. (1952). *An analysis of biographical inventory and Spatial Apperception Test scores in relation to other selection tests* (Special Report No. 52–5). Pensacola, FL: U.S. Naval School of Aviation Medicine.

Arth, T. O., Steuck, K. W., Sorrentino, C. T., & Burke, E. F. (1988) *Air Force Officer Qualifying Test (AFOQT): Predictors of undergraduate pilot training and undergraduate navigator training success* (Tech. Rep. No. AFHRL–TP–88–27). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Bair, J. T., Lockman, R. F., & Martoccia, C. T. (1956).Validity and factor analysis of naval air training predictor and criterion measures. *Journal of Applied Psychology, 40,* 213–219.

Bale, R. M., & Ambler, R. K. (1971). Application of college and flight background questionnaires as supplementary noncognitive measures for use in the selection of student naval aviators. *Aerospace Medicine, 42,* 1178–1181.

Bartram, D., & Dale, H. C. A. (1982). The Eysenck Personality Inventory as a selection test for military pilots. *Journal of Occupational Psychology, 55,* 287–296.

Berkshire, J. R. (1967). *Evaluation of several experimental aviation selection tests.* Pensacola, FL: Naval Aerospace Medical Center.

Berkshire, J. R., & Ambler, R. K. (1963). The value of indoctrination flights in the screening and training of Naval aviators. *Aerospace Medicine, 34,* 420–423.

Bordelon, V. P., & Kantor, J. E. (1986). *Utilization of psychomotor screening for USAF pilot candidates: Independent and integrated selection methodologies* (Tech. Rep. No. AFHRL–TR–86–4). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Burke, E. F. (1980). *Results of a preliminary study on a new tracking test for pilot selection* (Note No. 9/80). London: Ministry of Defence, Science 3 (Royal Air Force).

Carretta, T. R. (1987). The Basic Attributes Tests: An experimental selection and classification instrument for U.S. Air Force pilot candidates. In R. S. Jensen (Ed.), *Proceedings of the Fourth International Symposium on Aviation Psychology* (pp. 500–507). Columbus: Ohio State University, Aviation Psychology Laboratory.

Carretta, T. R., & Siem, F. M. (1988). *Personality, attitudes, and pilot training performance: Final analysis* (Tech. Rep. No. AFHRL–TP–88–23). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Cox, R. H. (1988). Utilization of psychomotor screening for USAF pilot candidates: Enhancing predictive validity. *Aviation, Space & Environmental Medicine, 59,* 640–645.

Croll, P. R., Mullins, C. J., & Weeks, J. L. (1973). *Validation of the cross-cultural aircrew aptitude battery on a Vietnamese pilot trainee sample* (Tech. Rep. No. AFHRL–TR–73–30). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Personnel Research Division.

Damos, D. L., & Lintern, G. (1979). *A comparison of single- and dual-task measures to predict pilot performance* (Engineering Psychology Tech. Rep. No. 79/2). Urbana–Champaign: University of Illinois at Urbana–Champaign.

Davis, R. A. (1989). *Personality: Its use in selecting candidates for U.S. Air Force undergraduate pilot training* (Research Rep. No. AU–ARI–88–8). Maxwell Air Force Base, AL: Air University Press.

DeWet, D. R. (1963). The roundabout: A rotary pursuit-test, and its investigation on prospective air-pilots. *Psychologia Africana, 10,* 48–62.

Doll, R. E. (1962). *Officer peer ratings as a predictor of failure to complete flight training* (Special Report No. 62–2). Pensacola, FL: U. S. Naval Aviation Medical Center

Elshaw, C. C., & Lidderdale, I. G. (1982). *Flying selection in the Royal Air Force. Revue de Psychologie Appliquie,* 32(Suppl.), 3–13.

Fiske, D. W. (1947). Validation of naval aviation cadet selection tests against training criteria. *Journal of Applied Psychology, 31,* 601–614.

Flanagan, J. C. (1947). *The aviation psychology program in the Army Air Forces* (Army Air Forces Aviation Psychology Program Research Report No. 1). Washington, DC: U.S. Government Printing Office.

Fleischman, H. L., Ambler, R. K., Peterson, F. E., & Lane, N. E. (1966). *The relationship of five personality scales to success in naval aviation training* (NAMI Rep. No. 968). Pensacola, FL: U.S. Naval Aviation Medical Center.

Fleishman, E. A. (1954). *Evaluations of psychomotor tests for pilot selection: The direction control and compensatory balance tests* (Tech. Rep. No. AFPTRC–TR–54–131). Lackland Air Force Base, TX: Air Force Personnel and Training Research Center.

Fleishman, E. A. (1956). Psychomotor selection tests: Research and application in the United States Air Force. *Personnel Psychology, 9,* 449–467.

Flyer, E. S., & Bigbee, L. R. (1954). *The light plane as a pre-primary selection and training device: III. Analysis of selection data* (Tech Rep. No. AFPTRC–TR–54–125). Lackland Air Force Base, TX: Air Force Personnel and Training Research Center.

Fowler, B. (1981). The aircraft landing test: An information processing approach to pilot selection. *Human Factors, 23,* 129–137.

Goebel, R. A., Baum, D. R., & Hagin, W. V. (1971). *Using a ground trainer in a job sample approach to predicting pilot performance* (Tech. Rep. No. AFHRL–TR–71–50). Williams Air Force Base, AZ: Air Force Human Resources Laboratory, Flying Training Division.

Gopher, D. (1982). A selective attention test as a predictor of success in flight training. *Human Factors, 24,* 173–183.

Gopher, D., & Kahneman, D. (1971). Individual differences in attention and the prediction of flight criteria. *Perceptual and Motor Skills, 33,* 1335–1342.

Gordon T. (1949). The airline pilot's jobs. *Journal of Applied Psychology, 33,* 122–131.

Graybiel, A., & West, H. (1945). The relationship between physical fitness and success in training of U.S. Naval flight students. *Journal of Aviation Medicine, 16,* 242–249.

Greene, R. R. (1947). Studies in pilot selection: II. The ability to perceive and react differentially to configuration changes as related to the piloting of light aircraft. *Psychological Monographs, 61,* 18–28.

Griffin, G. R., & McBride, D. K. (1986). *Multitask performance: predicting success in naval aviation primary flight training* (Tech. Rep. No. NAMRL–1316). Pensacola, FL: U.S. Naval Aerospace Medical Research Laboratory.

Griffin, G. R., & Mosko, J. D. (1982). *Preliminary evaluation of two dichotic listening tasks as predictors of performance in naval aviation undergraduate pilot training* (Tech. Rep. No. NAMRL–1287). Pensacola, FL: U.S. Naval Aerospace Medical Research Laboratory.

Guilford, J. P., & Lacey, J. I. (1947). *Printed classification tests* (Army Air Forces Aviation Psychology Program Research Report No. 5). Washington, DC: U.S. Government Printing Office.

Guinn, N., Vitola, B. M., & Leisey, S. A. (1976). *Background and interest measures as predictors of success in undergraduate pilot training* (Tech. Rep. No. AFHRL–TR–76–9). Lackland Air Force Base, TX: Air Force Human Resources Laboratory, Personnel Research Division.

Hertli, P. (1982). *The prediction of success in Army aviator training: A study of the warrant officer candidate selection process.* Unpublished manuscript, U.S. Army Research Institute Field Unit, Fort Rucker, AL.

Hunter, D. R. (1982). *Air Force pilot selection research.* Paper presented at the 90th meeting of the American Psychological Association, Washington, DC.

Hunter, D. R., & Thompson, N. A. (1978). *Pilot selection system development* (Tech. Rep. No. AFHRL-TR-78-33). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Personnel Research Division.

Joaquin, J. B. (1980). *The Personality Research Form (PRF) and its utility in predicting undergraduate pilot training performance in the Canadian Forces* (Working Paper No. 80–12). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

Kaplan, H. (1965). *Prediction of success in Army aviation training* (Tech. Rep. No. 1142). Washington, DC: U.S. Army Personnel Research Office.

King, J. E. (1945). Relation of aptitude tests to success of Negro trainees in elementary pilot training (Research Bulletin No. 45-52). Tuskegee Army Air Field, AL: Office of the Surgeon, Headquarters Army Air Forces Training Command.

Knight, S. (1978). *Validation of RAF pilot selection measures* (Note No. 7/78). London: Ministry of Defence, Science 3 (Royal Air Force).

Koonce, J. M. (1981). Validation of a proposed pilot trainee selection system. In R. S. Jensen

(Ed.), *Proceedings of the First Symposium on Aviation Psychology* (Tech. Rep. No. APL–1–81, pp. 255–260). Columbus: Ohio State University, Aviation Psychology Laboratory.

Lane, G. G. (1947). Studies in pilot selection: I. The prediction of success in learning to fly light aircraft. *Psychological Monographs, 61,* 1–17.

LeMaster, W. D., & Gray, T. H. (1974). *Ground training devices in job sample approach to UPT selection and screening* (Tech. Rep. No. AFHRL–TR–74–86). Williams Air Force Base, AZ: Air Force Human Resources Laboratory, Flying Training Division.

Lidderdale, I. G. (1976). *The primary flying grading trial interim report No. 2.* RAF Brampton, England: Ministry of Defence, Royal Air Force, Headquarters Command, Research Branch.

McAnulty, D. M. (1990). *Validation of an experimental battery of Army aviator ability tests.* Fort Rucker, AL: Anacapa Sciences, Inc.

McGrevy, D. F., & Valentine, L. D. (1974). *Validation of two aircrew psychomotor tests* (Tech. Rep. No. AFHRL–TR–74–4). Lackland Air Force Base, TX: Air Force Human Resources Laboratory, Personnel Research Division.

Melton, A. W. (Ed.). (1947). *Apparatus tests* (Army Air Forces Aviation Psychology Research Report No. 4). Washington, DC: U.S. Government Printing Office.

Miller, J. T., Eschenbrenner, A. J., Marco, R. A., & Dohme, J. A. (1981). *Mission track selection process for the Army initial entry rotary wing flight training program.* St. Louis, MO: McDonnell Douglas.

Mullins, C. J., Keeth, J. B., & Riederich, L. D. (1968). *Selection of foreign students for training in the United States Air Force* (Tech. Rep. No. AFHRL–TR–68–111). Lackland Air Force Base, TX: Air Force Human Resources Laboratory, Personnel Research Division.

North, R. A., & Gopher, D. (1974). Basic attention measures as predictors of success in flight training (Tech. Rep. No. ARL–74–14). Urbana–Champaign: University of Illinois at Urbana–Champaign, Aviation Research Laboratory.

Owens, J. M., & Goodman, L. S. (1983). *Navy aviation selection and classification research.* Paper presented at the 11th meeting of the Department of Defense Human Factors Engineering Technical Advisory Group, Aberdeen Proving Grounds, MD.

Roth, J. T. (1980). *Continuation of data collection on causes of attrition in initial entry rotary wing training.* Valencia, PA: Applied Science Associates.

Sells, S. B. (1956). Further developments on adaptability screening of flying personnel. *Journal of Aviation Medicine, 27,* 440–451.

Sells, S. B., Trites, D. K., Templeton, R. C., & Seaquist, M. R. (1958). Adaptability screening of flying personnel: Cross validation of the personal history blank under field conditions. *Journal of Aviation Medicine, 29,* 683–689.

Shipley, B. D. (1983). *Maintenance of level flight in a UH-1 flight simulator as a predictor of success in Army flight training.* Unpublished manuscript, U.S. Army Research Institute, Fort Rucker, AL.

Shoenberger, R. W., Wherry, R. J., & Berkshire, J. R. (1963). *Predicting success in aviation training* (Report No. 7). Pensacola, FL: U.S. Naval Aviation Medical Center.

Shull, R. N., & Dolgin, D. L. (1989). Personality and flight training performance. In *Proceedings of the 33rd Annual Meeting of the Human Factors Society.* Santa Monica, CA: Human Factors Society.

Shull, R. N., Dolgin, D. L., & Gibb, G. D. (1988). *The relationship between flight training performance, a risk assessment test, and the Jenkins Activity Survey* (Tech. Rep. No. NAMRL–1339). Pensacola, FL: U.S. Naval Aerospace Medical Research Laboratory.

Siem, F. M. (1988). Personality characteristics of USAF pilot candidates. In *Human behaviour in high stress situations in aerospace operations* (AGARD Conference Proceedings No. 458, pp. 6-1–6-7). Paris: North Atlantic Treaty Organization, Advisory Group for Aerospace Research and Development.

Signori, E. I. (1949). The Arnprior experiment: A study of World War II pilot selection procedures in the RCAF and RAF. *Canadian Journal of Psychology, 3,* 136–150.

Stoker, P. (1982). An empirical investigation of the predictive validity of the defence mechanism test in the screening of fast-jet pilots for the Royal Air Force. *Projective Psychology, 27,* 7–12.

Stoker, P., Hunter, D. R., Kantor, J. E., Quebe, J. C., & Siem, F. M. (1987). *Flight screening program effects on attrition in undergraduate pilot training* (Tech. Rep. No. AFHRL–TP–86–59). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Trankell, A. (1959). The psychologist as an instrument of prediction. *Journal of Applied Psychology, 43,* 170–175.

Tucker, J. A. (1954). *Use of previous flying experience as a predictor variable* (Tech. Rep. No. AFPTRC–TR–54–71). Lackland Air Force Base, TX: Air Force Personnel and Training Research Center.

Voas, R. B. (1959). Vocational interests of naval aviation cadets: Final results. *Journal of Applied Psychology, 43,* 70–73.

Want, R. L. (1962). The validity of tests in the selection of Air Force pilots. *Australian Journal of Psychology, 14,* 133–139.