AFRL-HE-WP-TR-1999-0232

UNITED STATES AIR FORCE RESEARCH LABORATORY

PITFALLS OF ABILITY RESEARCH IN AVIATION PSYCHOLOGY

Thomas R. Carretta

HUMAN EFFECTIVENESS DIRECTORATE
CREW SYSTEM INTERFACE DIVISION
WRIGHT-PATTERSON AFB OH 45433-7022

Malcolm James Ree

CENTER FOR LEADERSHIP STUDIES
OUR LADY OF THE LAKE UNIVERSITY
411 SW 24th STREET
SAN ANTONIO TX 78207

DECEMBER 1999

INTERIM REPORT FOR THE PERIOD SEPTEMBER 1998 TO JANUARY 1999

Human Effectiveness Directorate
Crew System Interface Division
2255 H Street
Wright-Patterson AFB OH 45433-7022

NOTICES

Using Government drawings, specifications, or other data included in this document for any purpose other than Government-related procurement does not in any way obligate the US Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation, or convey any right or permission to manufacture, use, or sell any patented invention that may relate to them.

Please do not request copies of this report from the Air Force Research Laboratory. Additional copies may be purchased from:

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161

Federal Government agencies registered with the Defense Technical Information Center should direct request for copies of this report to:

Defense Technical Information center
8725 John J. Kingman Road, Suite 0944
Ft. Belvoir, VA 22060-6218

TECHNICAL REVIEW AND APPROVAL

AFRL-HE-WP-TR-1999-0232

The Office of Public Affairs has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This Technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

MARIS M. VIKMANIS
Chief, Crew Systems Interface Division
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 1999	Interim - September 1998 to January 1999

4. TITLE AND SUBTITLE

Pitfalls of Ability Research in Aviation Psychology

5. FUNDING NUMBERS

PE - 62202F
PR - 1123
TA - B1
WU - 01

6. AUTHOR(S)

Thomas R. Carretta*
Malcolm James Ree**

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory*
Human Effectiveness Directorate
Crew System Interface Division
Air Force Materiel Command
Wright-Patterson AFB OH 45433-7022

Center for Leadership Studies**
Our Lady of the Lake University
411 SW 24th Street
San Antonio TX 78207

8. PERFORMING ORGANIZATION REPORT NUMBER

AFRL-HE-WP-TR-1999-0232

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
Human Effectiveness Directorate
Crew System Interface Division
Air Force Materiae Command
Wright-Patterson AFB OH 45433-7022

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

Air Force Research Laboratory Technical Monitor: Dr. Thomas R. Carretta, (937) 656-7014; DSN 986-7014

12a. DISTRIBUTION AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Ability research in aviation psychology can be fraught with pitfalls that lead to inappropriate conclusions. We identify several issues that lead to potential misinterpretation of results and suggest corrective solutions. These issues include lack of construct validity of the measures, misinterpretation of correlations and regression weights, lack of statistical power, failure to estimate cross-validation effects, and misinterpretation of factor analytic results.

14. SUBJECT TERMS

Methodological issues
Unreliability
Range Restriction
Interpretation of correlations
Statistical power
Factor analysis
Cross-validation

15. NUMBER OF PAGES

21

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

This page intentionally left blank.

PREFACE

This effort was performed under work unit 1123-B1-01 in support of USAF aircrew selection and classification. Send e-mail for Thomas R. Carretta to thomas.carretta@he.wpafb.af.mil. Send e-mail for Malcolm James Ree to MREE@STIC.NET.

TABLE OF CONTENTS

FIGURES

PITFALLS OF ABILITY RESEARCH IN AVIATION PSYCHOLOGY

SUMMARY

Ability research in aviation psychology can be fraught with pitfalls that lead to inappropriate conclusions. We identify several issues that lead to potential misinterpretation of results and suggest corrective solutions. These issues include lack of construct validity of the measures, misinterpretation of correlations and regression weights, lack of statistical power, failure to estimate cross-validation effects, and misinterpretation of factor analytic results.

INTRODUCTION

Selection of applicants for pilot training or other aviation employment frequently is based on theories about the relationship between ability and job performance. However, several methodological issues (Carroll, 1978; Ree, 1995; Ree & Carretta, 1997) may affect the interpretability of the ability-job performance relationship. These issues fall into the following categories:

1. lack of construct validity of the measures
2. misinterpretation of correlations and regression weights
    a. holding job experience constant
    b. effects of range restriction
    c. effects of unreliability of measures
    d. effects of dichotomization of criteria
    e. examination of effects for subgroups
    f. weighting of variables
3. lack of statistical power
4. failure to estimate cross-validation effects
5. misinterpretation of factor analytic results

The purpose of this paper is to illustrate each of these issues for aviation psychology, explain the consequences of failing to address each of them correctly, and suggest solutions to the problems caused by the issues.

METHODOLOGICAL ISSUES

Misunderstanding Construct Validity

Construct is the term used to describe an abstraction such as "intelligence," "workload," "situation awareness," or "the right stuff." Constructs are not directly observable and must be inferred from some measurement scale or test that operationalizes the elements of the construct. A construct that cannot be operationally defined and measured has no scientific value.

There is a tendency for researchers to engage in the "topological fallacy" (Walters, Miller, & Ree, 1993), that is to interpret a measurement scale or test on the basis of the topology or

appearance of the questions. The problem is that the appearance of the questions is not necessarily a true indicator of the construct measured by the scale or test. The only way to know what a scale or test measures is to administer the scale or test with other scales or tests that are agreed upon measures of the construct or constructs of interest. This procedure is often called "administering the test in the presence of a reference battery" (see French, Ekstrom, & Price, 1969 for an example). To know what your scale or test measures, administer it along with accepted measures of the construct.

Two examples illustrate this point well. Rabbitt, Banerji, and Szymanski (1989) estimated the correlation between an ordinary IQ test (AH 4; Heim & Batts, 1948) and a complex computerized task called "Space Fortress" (Donchin, Fabiani, & Sanders, 1989). On the surface (topology), the IQ test and Space Fortress look very different. The paper-and-pencil IQ test has the familiar form of questions and answers. In contrast, Space Fortress requires monitoring a computer screen with many moving icons, learning several complex rules, and manipulating control sticks and foot pedals and pushing buttons. In most instances where Space Fortress has been used, it is common that participants practice the test over several hours, separated by breaks between sessions. Despite the apparent differences between the IQ test and Space Fortress, Rabbitt et al. found a correlation between them of .69 (uncorrected), a value as high as normally found between two different IQ tests.

Walters et al. (1993) who examined the validity and incremental validity of an experimental structured interview for selecting US Air Force pilot candidates provided another example. Two hundred twenty-three (223) participants completed the Air Force Officer Qualifying Test (AFOQT; Carretta & Ree, 1996), experimental computer-based cognitive and personality tests, and the structured interview. The interview was conducted by experienced Air Force pilots who had completed a course in interview techniques. The interview included questions regarding educational background, flying job knowledge, motivation to fly, and self-confidence and leadership. In addition to rating the pilot candidates in these 4 areas, interviewers also rated them on probability of success in pilot training, bomber/fighter training, and tanker/transport training. The dependent measure in the validation study was passing/failing pilot training. Even though the 7 interview scores demonstrated validity against pilot training outcome, they failed to provide incremental validity when used along with the AFOQT and computer-based test scores. Subsequent to Walters et al., linear regression analyses were done to evaluate full and restricted regression models for predicting pilot training outcome. The predictors were a measure of general cognitive ability (*g*) extracted from the AFOQT and the 7 interview scores. These analyses showed that adding the interview scores to the measure of *g* did not improve prediction. The lack of incremental validity for the interview scores occurred because they lacked unique predictive variance. The predictive utility of the interview clearly came from its measurement of *g*.

These two studies demonstrated the value of using a reference test. Rabbitt et al. (1989) used the AH 4 as a reference test and Walters et al. (1993) used the AFOQT as a reference test. Had the researchers not used reference tests they might have concluded that their measures were different from existing constructs. Because they used reference tests they understood their results and were able to interpret the factors measured by Space Fortress and the structured interview.

Misinterpreting Correlations

Holding Job Experience Constant

Ability research is generally correlational in nature. The interpretation of correlations, although straightforward on the surface, can be fraught with hazards. Consider the correlation of an ability test and ratings of pilot job performance. It would be usual to find low correlations which could lead to inappropriately abandoning predictive measures. The *Principals for the Validation and Use of Personnel Selection Procedures* (APA, 1987) notes that the relationship between ability (or any other measure) and occupational criteria is best understood with the effect of job experience removed. This can easily be done by using partial correlation and "partialing-out" experience from the relationship between ability and the criteria. Carretta, Perry, and Ree (1996) provide an example. They correlated ability test scores with ratings of situational awareness (SA) for 171 F-15 pilots. The zero-order correlation (zero-order is the term used to indicate that no partialling-out has been done) of ability and SA was .10. However, when F-15 flying experience was partialed-out, the correlation was .17. In this instance, it would be incorrect to report the correlation of ability and SA as .10.

More broadly, the idea of partial correlation can be subsumed under mediation. Mediation means that one variable acts through another to exert its influence on a third variable. For example, "A →B→ C" indicates that variable A acts through variable B to exert its influence on variable C. Note that there is no direct influence of A on C. We do not specify "A→C." This should not be interpreted to mean that variable A has no influence on variable C, but rather that A works through B to influence C. Hunter (1986) has provided an informative model of mediation in the area of job performance. In numerous jobs, he demonstrated that job knowledge mediated the relationship between ability and job performance. Ree, Carretta, and Teachout (1995) demonstrated this mediation for pilot trainees. In the Ree et al. study, *g* had both direct and indirect influence on the acquisition of aviation job knowledge and hands-on flying performance during pilot training. To know the true relationship of a predictor to job performance, it is necessary to partial-out the effect of job experience.

Adverse Effects of Range Restriction

Censored samples are encountered frequently in studies of human performance. Censoring occurs when the variance of one or more variables has been reduced due to prior selection. This reduction in variance is usually referred to as range restriction. For example, employers typically do not hire all job applicants, nor do universities admit all those who apply. A common example of a censored sample is provided by job applicants who have already been screened on the basis of educational achievement (e.g., completion of a college degree in an appropriate specialty, selection interview) and vocational interest. Similarly, university students have been subjected to prior selection on the basis of standardized test scores and prior academic achievement. As a result of the censoring, the variances associated with the causes of academic achievement, job performance, or test scores have been reduced. Censored samples have been shown to create artifacts that may lead to erroneous conclusions (Morrison & Morrison, 1995) which could lead to inappropriately abandoning predictive measures.

Censored samples can produce estimates of correlations that are substantially downwardly biased (Martinussen, 1997; Thorndike, 1949). Correlations based on censored samples can sometimes even change signs from their population value (Ree, Carretta, Earles, & Albert, 1994; Thorndike, 1949). Despite claims to the contrary, when computed in a censored sample the correlation between a personnel selection test and a performance criterion will be much lower than it should be.

Thorndike (1949, pp. 170-171) provided a dramatic illustration of the problems that can occur when censoring exists. During WWII, an experimental group of 1,036 US Army Air Force aircraft pilot applicants was admitted to training without regard to their scores on 5 aptitude tests. Subsequently, correlations with the training criterion were computed for all participants (n = 1,036) and for those pilot candidates (n = 136 out of 1,036) that would have been selected had the strict standards in effect at the end of WWII been used. Compared to the unrestricted sample, the average decrease in the 5 validity coefficients was .29 in the sample of 136 qualified pilot candidates (i.e., the range-restricted or censored sample). In the unrestricted sample, the Pilot Stanine composite derived from the 5 tests had a correlation of .64 with training outcome. It dropped to .18 in the range-restricted sample. The most dramatic change occurred for a psychomotor test where the correlation changed sign from +.40 to -.03 from the unrestricted to the range-restricted sample. It is clear that the validity estimates were adversely affected by range restriction in this case. Further, had only the range-restricted correlations been reported, wrong decisions would have been made as to which tests to implement.

Range restriction appears to be commonplace in psychological research. Brand (1987) noted that many studies have been conducted in samples that were range restricted on general cognitive ability (*g*). It is not unusual to find studies on the predictiveness of aptitude tests reported in the open literature where the participants (e.g., college or university students, military enlistees, job incumbents) were preselected on the basis of ability or prior experience. This reduces the variance of ability and artificially reduces correlations. Goldberg (1991) noted that "in other words, one can always filter out or at least greatly reduce the importance of a causal variable, no matter how strong that variable, by selecting a group that selects its members on the basis of that variable" (p. 132). He observed that the more restricted the variance of a variable, the less its apparent validity. In other words, the greater the range restriction imposed on the variable by prior selection the less validity it will seem to have.

Statistical corrections are available and should be applied to reduce problems resulting from range restriction and thus provide better statistical estimates. The "univariate" corrections described by Thorndike (1949) are appropriate if censoring has occurred on only one variable. However, if censoring has occurred on more than one variable, the multivariate correction (Lawley, 1943; Ree et al., 1994) is more appropriate. The corrections described by Thorndike (1949) and Lawley (1943) provide better statistical estimates and tend to be conservative (Linn, Harnish, & Dunbar, 1981). That is, the corrections still tend to underestimate the population values. Johnson and Ree (1994) offered a free windows-based software program that can perform either univariate or multivariate corrections. Correction for range restriction should always be used when working with range restricted samples.

Reduction from Unreliability of Measures

The use of unreliable measures reduces correlations (Spearman, 1904) and can lead to incorrect conclusions. The magnitude of the correlation between two variables is limited by a function of the reliability of the two variables. Correcting for unreliability (also called "the correction for attenuation") in validation studies informs us about the nature of the true relationship between predictors and criteria. Correlations between tests and a criterion that change from low to moderate or high after correction suggest that the test could help increment validity if it (or the criterion) were more reliable. If validities remain low to moderate after correction for unreliability this would suggest that the criterion has other sources of variance that are not being predicted.

Carretta and Ree (1995) provided an example. They examined the validity of the 16 AFOQT tests against US Air Force undergraduate pilot training grades, daily flying grades, and check flight grades. The AFOQT validities were examined as observed, corrected for range restriction, and fully-corrected for both range restriction and unreliability (of the predictor and criterion). The average magnitude of the correlations of the 16 tests with the criteria varied from .0268 to .1332 for the observed correlations, from .0986 to .2463 for the range-restriction-corrected correlations, and from .2498 to .5792 for the fully-corrected correlations. Clearly, use of the appropriate corrections removes statistical artifacts that are inevitable in all studies (Hunter & Schmidt, 1990).

Although not as widely known, unreliability has an effect on regression coefficients. Fuller (1987) demonstrated mathematically that the b-weight estimate of β is biased by unreliability. Instead of b estimating β, it estimates $r_{xx'}\beta$ (reliability times β) reducing the slope of the regression line. Similarly, the regression constant is affected and biased by unreliability. Jensen (1980, p. 463) provides a good discussion of the problem and equations to correct regression b-weights and constants for the effects of unreliability. Carretta (1997) applied Jensen's equations to differing regression intercepts computed in a study of predictive bias for US Air Force pilot trainees and found no difference in intercepts after correction. Carretta observed that "The uncritical interpretation of different intercepts...is unwarranted" (p. 125).

Unreliability of measures also plays a part in factor analysis whether exploratory or confirmatory. Just as reliability attenuates correlations and regression coefficients, it attenuates factor loadings. This causes the factor loadings to be underestimates of the true values. This underestimation can be corrected by dividing the factor loading by the reliability. Ree and Earles (1993) reported data from Jones (later published as Jones & Ree, 1998) that showed the correlation of factor loadings and validity for a series of Air Force jobs to be .78. Correcting the factor loadings for the unreliability of the tests, the correlation was estimated at .98.

To improve validity one solution is to add items to unreliable measures (e.g. more test questions, additional job performance ratings, etc.). Other solutions might be to remove ambiguity from existing items, improve instructions, and reduce ambiguity from scoring. If these remedies fail, it may be necessary to discard the measure.

In any case, reliability should be estimated for all tests and training or job performance criteria. Depending on the circumstances, different types of reliability estimates (i.e., internal consistency, test-retest) are appropriate. These reliability estimates can then be used in the correction for attenuation formulas (Spearman, 1904, Hunter & Schmidt, 1990) and the resulting corrected correlations can be used to address theoretical issues.

Path analysis (Agresti & Finlay, 1997) is another category that benefits from correction for unreliability. In a path analysis, an independent variable is said to cause or "explain" a dependent variable. As path coefficients are standardized regression coefficients, unreliability attenuates the estimates. Failure to correct for unreliability of the independent variable will necessarily lead to underestimation of causal effects.

Dichotomization of Criteria

When studying training success of pilots or other aviation jobs, it is not unusual to have criteria that fall into two categories, pass or fail. Did the student pilot pass or fail the course? Did the aviation mechanic pass or fail the airframe certification test? Dichotomization of criteria causes correlations to appear lower than they should. It places an upper limit on the magnitude of the correlation that depends on the proportion in each of the pass and fail categories. When the proportions are 50-50 there is little adverse effect on the correlation, but when the proportions deviate from 50-50, a downward bias effects the correlation. For example, if the correlation between two variables is about .5 before dichotomization and the proportions are 50-50 in the dichotomized criterion, the correlation found in the study will be .5. If the proportions are 60-40, 70-30, 80-20, and 90-10 the correlations will be .39, .38, .35, and .29, respectively. If the correlation before dichotomization were about .25, the after-dichotomization correlational values for the proportions 60-40, 70-30, 80-20, and 90-10 would be: .20, .19, .17, and .15. This has long been recognized as a problem. A statistical correction for the dichotomization (Cohen, 1983) provides an estimate of the correlation had the variable not been dichotomized.

Examination of Effects for Subgroups

In aircraft maintenance, it may be necessary to lift a heavy object above your head. During a validation study, a physical lifting test was administered. In these types of studies it is common to note that both sexes or two or more ethnic groups were included in the sample. Sometimes in the combined sex or ethnic groups the correlation between the predictor and criterion may be moderate or large, but within each group the correlation is low or zero. This means that the validity in the combined group may be nothing more than a statistical artifact. See for example Hogan (1991) and Ree, Carretta, and Earles (1999). This apparent paradox, moderate or high correlation in the combined group and low or zero correlation in each individual group, is not a psychological phenomenon, but a consequence of the mathematics of correlation and regression. Correlations and regression lines are data-driven. Figure 1 shows 3 examples where the predictor-criterion relationship is very different in two subgroups than in the combined group. In Figure 1a, there is a slight positive predictor-criterion relationship in each subgroup, but a much stronger positive relationship when the subgroups are combined. In Figure 1b, the correlation in each subgroup is negative, but is zero in the combined group. Finally, in Figure 1c, there is a slight negative correlation in one subgroup, a zero correlation in the other subgroup, but a strong

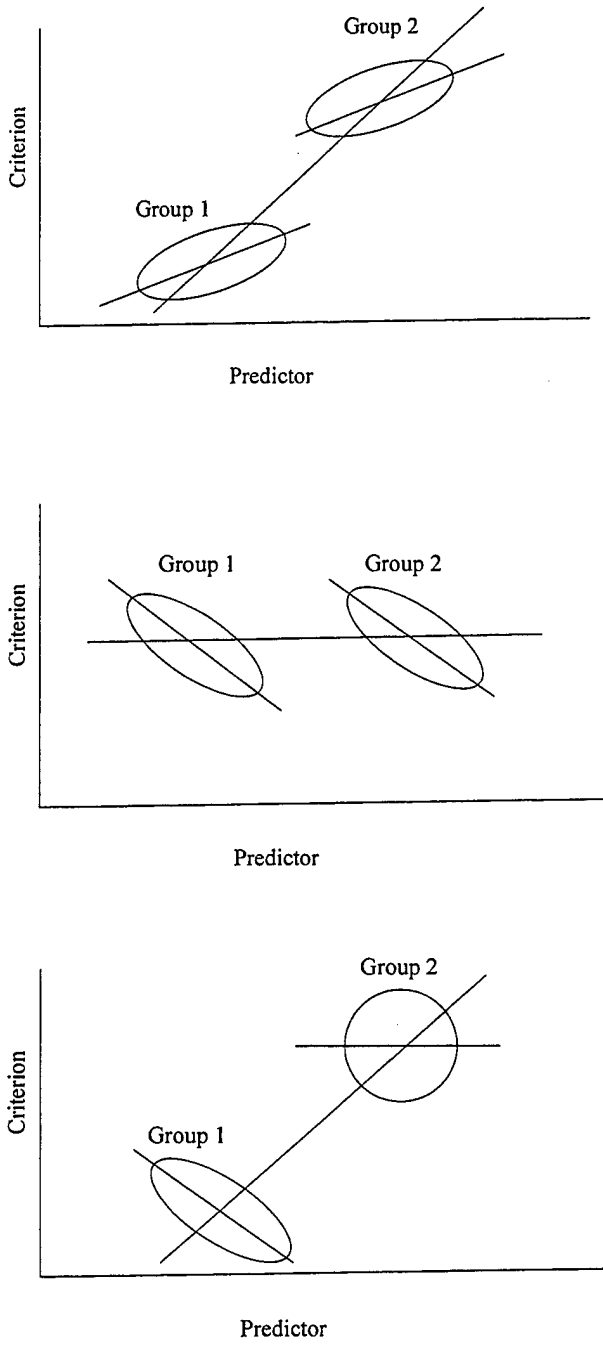

Figure 1. Examples of different predictor-criterion relationships for subgroups and combined group

Note. Figure 1a shows a weak positive predictor-criterion relationship for each subgroup and a stronger positive relationship for the combined group. Figure 1b shows a negative predictor-criterion relationship in each subgroup, but a zero relationship in the combined group. Figure 1c shows an example of a slight negative relationship in one subgroup, a zero relationship in the other subgroup, but a strong positive predictor-criterion relationship in the combined group.

positive correlation in the combined group. Indeed, it is possible to have nearly any combination of three correlations so long as the correlation in the combined group is neither +1 nor –1.

Ree, Carretta, and Earles (1999) provide and discuss several examples of this two-group phenomenon. Further, they propose and demonstrate a general hierarchical linear models analysis to address the issue. The first step in the Ree et al. approach is to test for equivalence of the variance errors of estimate (Gulliksen & Wilks, 1950; Jensen, 1980; Reynolds, 1982). If the errors of estimate are equal, a series of F tests of specified hierarchical linear models is appropriate. The first test in the hierarchical models series compares a linear model with two slopes and two intercepts with a model with only one slope and one intercept. A non-significant F indicates that there are no between-groups differences and analyses should be conducted at the within-group (combined groups) level. If a significant difference were found, additional tests of the differences between slopes and between intercepts would be performed. In addition to conducting these linear models analyses, it is recommended that practitioners plot their data for each subgroup and for the combined group.

This linear models approach is not optimum when comparing more than two groups. When there are more than two groups, a Within and Between Analysis (WABA; Dansereau, Alutto, & Yammarino, 1984).) is applicable.

Weighting of Variables

Typically, applicants for aviation jobs are given a series of tests, or several interviews, or a simulation task with many scores, or a combination of all of these. The selecting agency then has to make a decision and will frequently combine the scores and other applicant information by addition to form a composite. Sometimes the various parts of the composite will be given greater importance by weighting them more. The score on the composite, rather than its components, will be used to make a decision.

Weighting variables to create composites has been the subject of much empirical and analytical study. In aviation psychology, criterion-based regression-weighting frequently is used (Walters et al., 1993), even though several studies argue for unit or simple weighting. More than three decades ago, Aiken (1966) thought the controversy over the use of simple weights was settled and was surprised to find colleagues arguing for regression-based weights on an intuitive basis. A decade later Wainer (1976, 1978) showed small loss in predictive efficiency from equal weights when compared with regression weights. Wainer noted that selection usually involves ranking of applicants and top-down selection. Weighting schemes usually are of little importance for rank ordering and top-down selection.

Top-down decision making is common in many aircrew applications including personnel selection (e.g., estimating training suitability), job performance (e.g., ratings of situational awareness or decision-making ability), and organizational effectiveness (e.g., crew resource management). When top-down decisions are made, weighting variables does not matter because the rank ordering remains almost constant.

Wilks (1938) proved a general mathematical theorem showing that under common circumstances, almost all weighted composites of a set of variables are strongly correlated. In other words, if two different sets of weights were applied to a single set of variables to create two composites, the expected correlation between the two composites would be very high, frequently .99 or greater.

Ree, Carretta, and Earles (1998) demonstrated the effect of Wilks' (1938) theorem through several examples. They also provided several cases from published studies showing near identical rankings for composites based on unit weights, regression weights, policy capturing weights, and factor weights. Ree et al. showed that the magnitude of the correlation between two composites is a function of the coefficient of variation (CV; i.e., standard deviation of the weights divided by the mean of the weights), the number of variables in the composite, and the magnitude of the correlations among the variables. When weights are more variable (high CV), they affect rank-ordering on the composites only when the correlations among the variables are low. There will be little effect on rank-ordering due to the weights when the correlations are moderate or high. Correlations of human abilities relevant to aviation psychology such as situation awareness and crew resources management almost always show moderate to high correlations. Low correlations are not likely to be found in practice, except as artifacts due to range-restricted samples or unreliable measurement.

When considering weighting of variables it is appropriate to know which are important and then use simple or unit weights. Considering other than simple or unit weights, Wainer's (1976) said it succinctly "it don't make no nevermind."

Statistical Power

The probability of detecting a significant effect, such as a non-zero correlation or a difference between means when it is present, is called statistical power. More formally, it is the probability of rejecting the null hypothesis when it is false (Cohen, 1987). Although almost all statistics courses include the topic of statistical power, only a few published studies report power for the test statistics (such as t, Z, or F) being used (see for example Ree & Earles, 1991, 1993; Walters et al., 1993). Two surveys of a prestigious applied psychology journal showed that the average statistical power for studies accepted for publication was only .46 and declined to .37 two decades later (Sedlmeier & Gigerenzer, 1989). This means that the researchers could only expect to find an existing effect 46 percent of the time (or 37 percent) when it is there. Conversely, 54 percent of the time (or 63 percent off the time) they can expect to fail to find the existing effect! Why would a researcher conduct a study with less that a high chance of detecting a significant result? Low statistical power means that we are inclined to make incorrect conclusions about the psychological phenomena studied. Many aviation psychology studies have been conducted with small low-statistical power samples.

As noted elsewhere (Cohen, 1987), statistical power is a joint function of Type I error rate, effect size, sample size, and the degree to which the sample values reflect their true values in the population (i.e., the reliability of the sample values). Adjusting any one or all of these factors will affect the power of an analysis (i.e., the probability of rejecting the null hypothesis when it is false). Prior to conducting a study, Cohen (1987) should be consulted. Schmidt and Hunter

(1978) provide an informative discussion of sample size requirements (see especially page 222) and Schmidt, Hunter and Urry (1976) provide tables showing sample size requirements to achieve sufficient statistical power in validation studies given varying selection ratios, reliabilities, and effect sizes.

Cross-Validation

Multiple correlations, R, on samples are upwardly biased estimators of their population parameters. This means that the multiple correlation observed in a sample can be expected to decrease when the weights are applied to another sample. Wherry (1975) termed this phenomenon "45 years of shrinkage from overfitting." Mosier (1951) provided the classic paradigm for empirical cross-validation in which a single sample is drawn from a population and then divided into a validation and cross-validation sample. Regression weights are estimated in the validation sample and then applied in the cross-validation sample. Murphy (1983) has pointed out that there is only one sampling and that the estimated multiple correlation in the cross-validation sample is still a consequence of "overfitting" the data. Equally important is the fact that even if there were two samplings from the population, the validation and cross-validation multiple correlations would be only two values out of a virtually infinitely large set of values.

Further, the two-sample cross-validation approach is paradoxical and inconsistent. The goals of estimating regression weights are 1) stability of the estimate and 2) generalizability of the estimated parameters. Weights estimated in two half-samples of n_1 and n_2 cannot be as accurately estimated as from the entire sample of $N = n_1 + n_2$. As is well known, the standard error of a regression weight is a function of the sample size. Splitting a single sample into two pieces reduces the sample size used for estimation and increases the standard error. This reduces the accuracy of the estimate. However, the estimation of regression weights in only one sample provides no estimate of the cross-applicability (i.e., generalizability) of the weights.

Several non-sampling methods of estimating cross-validation shrinkage exist (Kennedy, 1988). Kennedy demonstrated the accuracy of an equation by Stein (1960) for estimating the mean of the distribution of all possible cross-validation multiple correlations from the population from which the sample was selected. Stein's Operator as it is called, has the advantage of allowing estimation on the largest available sample while offering an estimate of average cross-validity. It avoids the paradox.

Interpretation of Factor Analyses

Spearman (1904, 1927) developed factor analysis specifically for the study of human abilities. Factor analysis is the name given to a group of techniques for determining the sources of variance in a correlation matirx or a variance-covariance matrix. The need for the technique grew out of the observation that all tests of ability were positively correlated with one another. Spearman termed this phenomenon "positive manifold." Through factor analysis Spearman was able to detect latent sources of variability that influenced the correlation of the tests.

One problem in factor analysis is the misinterpretation of rotated factors. Factor rotation was developed to help make factors easier to interpret. However, when factors are rotated, variance

associated with the first factor is distributed among the remaining factors (Carroll, 1978; Jensen, 1980; Ree & Carretta, 1997; Ree & Earles, 1993). In ability tests, the first unrotated factor is typically a measure of general cognitive ability (*g*). When rotation occurs, the variance associated with the first factor seems to disappear, but in reality, it has become the dominant source of variance in the now rotated factors. Thus, these other factors become *g* plus something else. However, it is usually the "something else" that determines the label for the factor, while the general component is not acknowledged.

Consider an example. Kass, Mitchell, Grafton, and Wing (1983) factor-analyzed the Armed Services Vocational Aptitude Battery (ASVAB) and reported that it measured 4 orthogonal factors that accounted for 93% of the total variance: verbal, speeded performance, quantitative, and technical knowledge. The issue is how much of the scores represent pure measures of these 4 factors. To determine this, the factors either should not be rotated (see Olea & Ree, 1994; Ree & Earles, 1991; Ree, Earles, & Teachout, 1994) or the factor solution must be "residualized" (Schmid & Leiman, 1957). Residualization is based on performing a hierarchical factor analysis. In a hierarchical factor analysis the factors are rotated to an oblique solution and allowed to correlate. These factor correlations are then submitted to another factor analysis. Residualization removes the effects of the higher-order factors from the lower-order factors. For example, Ree and Carretta (1994) conducted a hierarchical factor analysis of the ASVAB and reported that the higher-order factor, *g*, accounted for 63.8% of the total variance and speed, verbal/math, and technical knowledge factors accounted for 6.2%, 2.4%, and 7.7% of the total variance, respectively. The results presented by Kass et al. are misleading as they omit the single largest source of variance, the higher order factor.

A similar example is provided by a comparison of factor analyses performed by Skinner and Ree (1987) and by Carretta and Ree (1996). Skinner and Ree conducted exploratory factor analyses of the Air Force Officer Qualifying test (AFOQT) on a sample of 3,000 US Air Force officer commissioning candidates and reported a 5-factor solution. They used a principal factors analysis with communalities in the principal diagonal and an oblique rotation. The factors all correlated positively with one another with an average correlation of .36. Based on the correlations among the oblique factors, Skinner and Ree speculated that one or more higher-order factors might be appropriate for the AFOQT, but did not test any hierarchical models. Carretta and Ree subsequently conducted confirmatory factor analyses of the Skinner and Ree data and found that a hierarchical solution provided the best fit. Carretta and Ree's 5 lower-order factors replicated those reported by Skinner and Ree. The single higher-order factor was identified as *g* and accounted for about 41% of the total variance. Again, failure to seek a hierarchical factor mislead Skinner and Ree as to the sources of variance for the test battery.

A third example was provided in a study of the determinants of situational awareness (SA) in US Air Force F-15 pilots. The criterion in the study consisted of peer and supervisor ratings of SA. The SA measures were developed (Houck, Wittaker, & Kendall, 1991, 1993) from task analyses conducted by psychologists with the assistance of 7 experienced F-15 pilots who served as subject matter experts (SMEs). Each of the pilot SMEs had more than 1,000 fighter aircraft hours. These SMEs identified several tasks necessary for air combat success and SA. The resulting SA scale included 31 behavioral items representing several broad groups of SA tasks (general, tactical game plan, systems operation, communication, information interpretation, and

tactical employment) in addition to overall ratings of SA and fighter ability. The issue of the factors in the SA scale was addressed by Carretta, Perry, and Ree (1996). They conducted a principle components analysis on peer and supervisory ratings of the SA scale for 171 US Air Force F-15 pilots and reported that the first unrotated principal component accounted for 92.5% of the variance, strongly suggesting that a single composite captured the ratings. Despite the great amount of effort put into identifying various aspects of SA and writing items to measure them, the data suggested a unidimensional construct.

The problem of the disappearing first factor as a result of rotation can be avoided by residualized hierarchical factors (Carretta & Ree, 1996) or by not rotating and using unrotated principal components or unrotated principal factors (Ree & Earles, 1991). Residualization and avoiding rotation are the only methods to produce pure measures of general and specific abilities for use in comparative analyses.

RECOMMENDATIONS

In this final section we present a list of recommendations for each of the issues raised in the introduction. Ability research is fraught with pitfalls that can lead to incorrect inferences. To avoid these traps do the following.

1. Use reference tests to establish construct validity. The appearance of a test is an unsure indicator.
2. When interpreting correlations:
   a. Hold job experience constant.
   b. Evaluate the utility of mediators. Some variables may exert their influence on others both directly and indirectly through some mediating variable.
   c. Correct correlations for statistical artifacts such as range restriction, unreliability of measures, and dichotomous variables. Less biased statistical estimates are preferable.
   d. When applicable, examine effects for subgroups as well as the total group. Relationships observed in the total group may be radically different from those observed in subgroups.
   e. Consider simple or unit weighting schemes to express the relationships among related variables (e.g., aptitude scores, job performance ratings). In many common instances, simple or unit weighting schemes are as effective as more complex and sometimes costly procedures (e.g., weights derived through policy capturing exercises or statistical procedures).
3. Take steps to ensure sufficient statistical power. It is wasteful to do studies when you have a low probability of detecting the effect.
4. Estimate cross-validities using one of several non-sampling methods. These non-sampling methods have the advantage of allowing estimation on the largest available sample while offering an estimate of cross-validity.
5. Misinterpretation of factor analysis results is often the direct consequence of rotation. The problem of the disappearing first factor as a result of rotation can be avoided by residualized hierarchical factors, or by using unrotated principal components or unrotated principal factors

REFERENCES

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.), Upper Saddle River, NJ: Prentice-Hall, 624-629.

Aiken, L. R., Jr. (1966). Another look at weighting test items. *Journal of Educational Measurement, 3,* 183-185.

American Psychological Association Division of Industrial-Organizational Psychology (1987). *Principles for the validation and use of personnel selection procedures* (3rd ed.). Washington, DC: Author.

Brand, C. (1987). The importance of general intelligence. In S. Modgil & C. Modgil (Eds.), *Arthur Jensen: Consensus and controversy.* New York: Falmer Press.

Carretta, T. R. (1997). Group differences on US Air Force pilot selection tests. *International Journal of Selection and Assessment, 5,* 115-127.

Carretta, T. R., Perry, D. C., Jr., & Ree, M. J. (1996). Predicting situational awareness in F-15 pilots. *The International Journal of Aviation Psychology, 6,* 21-41.

Carretta, T. R., & Ree, M. J. (1995). Air Force Officer Qualifying Test Validity for predicting pilot training performance. *Journal of Business and Psychology, 9,* 379-388.

Carretta, T. R., & Ree, M. J. (1996). Factor structure of the Air Force Officer Qualifying Test: Analysis and comparison. *Military Psychology, 8,* 29-42.

Carroll, J. B. (1978). How shall we study individual differences in cognitive abilities? – Methodological and theoretical perspectives. *Intelligence, 2,* 87-115.

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

Cohen, J. (1987). *Statistical power analysis for the behavioral sciences* (revised edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dansereau, F., Alutto, J. A., & Yammarino, F. J. (1984). *Theory testing in organizational behavior: The variant approach.* Englewood Cliffs, NJ: Prentice-Hall.

Donchin, E., Fabiani, M., & Sanders, A. (Eds.). (1989). The learning strategies program: An examination of the strategies in skill acquisition [Special Issue]. *Acta Psychologica, 71.*

French, J. W., Ekstrom, R. B., & Price, L. A. (1969). *Manual for kit of reference tests for cognitive factors* (Revised). Princeton, NJ: Educational Testing Service.

Fuller, W. A. (1987). *Measurement error models.* New York: Wiley.

Goldberg, S. (1991). *When wish replaces thought.* Buffalo, NY: Prometheus.

Gulliksen, H., & Wilks, S. S. (1950). Regression tests for several samples. *Psychometrika, 15,* 91-114.

Heim, A. W., & Batts, V. (1948). Upward and downward selection in intelligence testing. *British Journal of Psychology, 39,* 22-29.

Hogan, J. C. (1991). Physical abilities. In M. N. Dunnette & L. M. Hough (Eds.). *Handbook of Industrial and Organizational Psychology* (2nd ed.) Vol. 2 (pp. 753-871). Palo Alto, CA: Consulting Psychologists Press.

Houck, M. R., Whitaker, L. A., & Kendall, R. R. (1991). *Behavioral taxonomy for air combat: F-15 defensive counter-air mission* (UDR-TR-91-147).Dayton, OH: University of Dayton Research Institute.

Houck, M. R., Whitaker, L. A., & Kendall, R. R. (1993). *An information processing classification of beyond visual range air intercepts* (AL/HR-TR-1993-0061). Williams AFB, AZ: Armstrong Laboratory, Human Resources Directorate.

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

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis.* Newbury Park, CA: Sage.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

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

Jones, G.E., & Ree, M. J. (1998). Aptitude test validity: No moderating effects due to job ability requirements. *Educational and Psychological Measurement, 58,*282-292.

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

Kennedy, E. (1988). Estimation of the squared cross-validity coefficients in the context of best subtest regression. *Applied Psychological Measurement, 12,* 231-237.

Lawley, D. N. (1943). *A note on Karl Pearson's selection formulae. Proceedings of the Royal Society of Edinburgh, Section A, 62 (Pt.1),* 28-30.

Linn, R. L., Harnish, D. L., & Dunbar, S. (1981). Corrections for range restriction: An empirical investigation of conditions resulting in conservative corrections. *Journal of Applied Psychology, 66,* 655-663.

Martinussen, M. (1997). Pilot selection and range restriction: A red herring or a real problem? *Proceedings of the Ninth International Symposium on Aviation Psychology,* Columbus, OH, 1314-1318.

Morrison, T., & Morrison, M. (1995). A meta-analytic assessment of the predictive validity of the quantitative and verbal composites of the Graduate Record Examination with graduate grade point average representing the criterion of graduate success. *Educational and Psychological Measurement, 55,* 309-316.

Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement, 11,* 5-11.

Murphy, K. R. (1983). Fooling yourself with cross-validation: Single-sample designs. *Personnel Psychology, 36,* 111-118.

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than *g*. *Journal of Applied Psychology, 79,* 845-851.

Rabbitt, P., Banerji, N., & Szymanski, A. (1989). Space Fortress as an IQ test? Predictions of learning and of practiced performance in a complex interactive video-game. *Acta Psychologica, 71,* 243-257.

Ree, M. J. (1995). Nine rules for doing ability research wrong. *The Industrial-Organizational Psychologist, 32,* 64-68.

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

Ree, M., J. & Carretta, T. R. (1997). What makes an aptitude test valid? In R. F. Dillon (Ed.). *Handbook on testing* (pp. 65-81). Westport, CT: Greenwood Press.

Ree, M. J., Carretta, T. R., & Earles, J. A. (1998). In top-down decisions, weighting variables does not matter: A consequence of Wilks' theorem. *Organizational Research Methods, 1,* 407-420.

Ree, M. J., Carretta, T. R., & Earles, E. A. (1999). In validation, sometimes two sexes are one too many: A tutorial. *Human Performance, 12,* 79-88.

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. (1994). Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology, 79,* 298-301.

Ree, M. J., Carretta, T. R., & Teachout, M. S. (1995). Role of ability and prior job knowledge in complex training performance. *Journal of Applied Psychology, 80,* 721-730.

Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than *g*. *Personnel Psychology, 44,* 327-332.

Ree, M. J., & Earles, J. A. (1993). *g* is to psychology what carbon is to chemistry: A reply to Sternberg and Wagner, McClelland, and Calfee. *Current Directions in Psychological Science, 2,* 11-12.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than *g*. *Journal of Applied Psychology, 79,* 518-524.

Reynolds, C. E. (1982). Methods for detecting construct and predictive bias. In R. A. Bork (Ed.) *Handbook of methods for detecting test bias* (pp. 199-277). Baltimore, MD: Johns Hopkins University Press.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 55-61.

Schmidt, F. L., & Hunter, J. E. (1978). Moderator research and the law of small numbers. *Personnel Psychology, 31,* 215-232.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61,* 473-485.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309-316.

Skinner, J., & Ree, M. J. (1987). *Air Force Officer Qualifying Test (AFOQT): Item and factor analysis of form O* (AFHRL-TR-86-68). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Spearman, C. (1904). "General intelligence" objectively defined and measured. *American Journal of Psychology, 15,* 201-293.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* New York: Macmillan.

Stein, C. (1960). Multiple regression. In I. Olkin et al. (Eds.), *Contributions to probability and statistics.* Stanford, CA: Stanford University Press.

Thorndike, R. L. (1949). *Personnel selection.* New York: Wiley.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83,* 213-217.

Wainer, H. (1978). On the sensitivity of regression and regressors. *Psychological Bulletin, 85,* 267-273.

Walters, L. C., Miller, M., & Ree, M. J. (1993). Structured interviews for pilot selection: No incremental validity. *The International Journal of Aviation Psychology, 3,* 25-38.

Wherry, R. J. (1975). Underprediction from overfitting: 45 years of shrinkage. *Personnel Psychology, 29,* 1-18.

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika, 3,* 23-40.