INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

CHARACTERISTICS OF BIODATA KEYS AS A FUNCTION OF SCALING METHOD, SAMPLE SIZE, AND CRITERION

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

WILLIAM L. FARMER
Norman, Oklahoma
2002

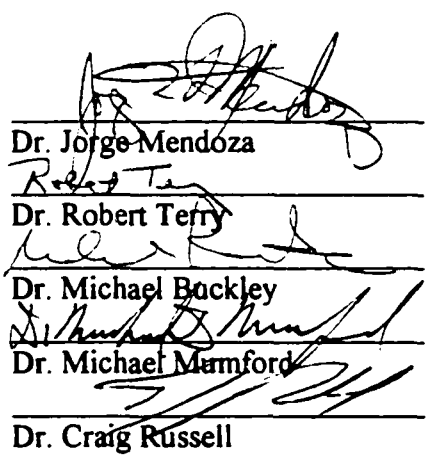
UMI Number: 3045829

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

CHARACTERISTICS OF BIODATA KEYS AS A FUNCTION OF SCALING METHOD, SAMPLE SIZE, AND CRITERION

A Dissertation APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY



Dr. Jorge Mendoza

Dr. Robert Terry

Dr. Michael Buckley

Dr. Michael Mumford

Dr. Craig Russell

Acknowledgements

I take this opportunity to thank all who have contributed in some way to this dissertation and to the completion of my graduate education. I first must thank my major professor and dissertation chair Dr. Jorge Mendoza. Jorge, you have been an excellent mentor and I truly appreciate your hanging in there with me for the duration of my time at OU. I have always valued your advice and friendship and cannot begin to place a value on the amount of influence that you've had on me. Again, thank you for everything.

From my committee I would like to thank Drs. Robert Terry, Mike Buckley, Mike Mumford, and Craig Russell. Dr. Terry, thank for serving as my dissertation co-chair and for the excellent guidance that I've received in measurement theory. Dr. Buckley, thank you for your support and for continually reminding me to never forget the "real world usefulness" and "bottom line" of my research. Dr. Mumford, thank you for your helpful insights into biodata research. Dr. Russell, your knowledge of biodata and your painstaking attention to detail have served as a model to me.

I would also like to thank Drs. Joe Rodgers and Larry Toothaker for providing a substantial part of the training that has brought me to this point. Dr. Rodgers, your approach to teaching and clarity of presentation were invaluable in getting me to understand topics that oftentimes were less than easy to fathom. Dr. Toothaker, you have served as a professional and personal model for me and if I can touch one person as positively as you have touched many, I will feel that I've accomplished much. I also take this opportunity to thank Nanette Bryant for all of the support I've received over the years, and the many others at OU who have played a part in my development.

As significant as my formal training at OU has been, I would be remiss in not acknowledging the major impact that working for the FAA Civil Aeromedical Institute has had on me professionally. Here I learned to use the tools I acquired in class and in the meantime developed into a research psychologist. I thank Drs. Bill Collins and Dave Schroeder for the financial and professional support that I received as I worked toward realizing my goals. I thank Drs. Tom Hilton and Edna Fiedler for the many years of management support and scientific insight. Dr. Fiedler, thank you also providing the necessary administrative approval to use FAA data for this project.

Others from the FAA who have contributed significantly are: Janine King, who provided more technical and logistical support than I could begin to adequatelv acknowledge (including the initial read through and editing of this document); Dr. Dana Broach, thank you for the many years of guidance and instilling in me the "selection bug;" Lendell "Coach" Nye, thank you for everything (Where do I begin?); and Nelda Milburn, thanks for the support and camaraderie as we walked parallel miles. To the others who've not been named, thank you for everything. Additionally, I owe a debt of gratitude to Mr. Murray Rowe and Dr. David Alderton, both of the Navy Personnel Research, Studies, and Technology Department of the Navy Personnel Command, for bearing with me as I completed this project.

On a more personal note, I truly appreciate all that the members of my family have done that have contributed so much to my education and career. My mother and father continue to exert a strong influence on me and I am always thankful for the guidance that they provided in my formative years. To my children, Katie and Collin, I give thanks for all of the resilience and patience that you have demonstrated for these

many years. It has not been easy for you to understand the importance of the sacrifices that you have made, but your "being yourselves" has made this journey more bearable and has helped me understand and appreciate the value of not being so serious all of the time.

Finally, and most importantly, I give an ocean of thanks to my wife Pam. Your support and patience throughout have provided the bedrock that I needed to stick with this and see it through to completion. I know that I can never replace all of the missed time together that this commitment has required, but I will always appreciate all that you have done. Maybe in some small way I can make it up to you. Your love has given me the strength that I needed. I dedicate this dissertation to you and the kids.

Table of Contents

List of Tables

List of Tables (continued)

List of Tables (continued)

List of Tables (continued)

List of Figures

Abstract

The present study investigated the impact of scaling method, key developmental sample size, and criterion characteristics on the incremental validity of scored biographical information (biodata) as a predictor of subsequent performance. Biodata surveys administered to air traffic control students at the FAA Academy between 1988 and 1990 were used to generate data that was scaled using three different measurement technologies: a) vertical percent empirical keying, b) factor analysis, and c) multidimensional item response theory (MIRT). The resulting biodata scale scores were combined with an ability-based selection test to predict FAA Academy performance (defined as final score and pass/fail) and air traffic field facility OJT performance (defined as certification or "wash out"). Comparisons were made between each method, for samples ranging in size from 4,568 to 100 students, regarding the amount of shrinkage between developmental and cross-validation groups, magnitudes of resulting cross-validities, and classification accuracy. The results indicated that at large sample sizes ($N \geq 1{,}000$) the differences between methods regarding the amount of shrinkage and the magnitude of cross-validity coefficients were negligible. At smaller sample sizes the amount of shrinkage increased with the empirical key enhanced model demonstrating a greater amount than the factor analytically derived keying model. Though differences were small, the incremental factor analytic model appeared to predict eventual performance, in most of the samples, better than the empirical or MIRT methods. This study contributes to the comparative scaling biodata literature by providing guidelines for future research and practice. Another outcome is exposure of MIRT methods to the biodata research community.

CHAPTER 1 - INTRODUCTION

The purpose of the present study was to investigate the impact scaling method, key developmental sample size, and criterion characteristics have on the incremental validity of biographical information (biodata) as a predictor of subsequent performance. The problem addressed was: does the amount of incremental variance accounted for by keyed biodata, over and above measures of ability, for predicting performance vary depending on how the data is scaled, the sample size used to develop the scale, and the distributional properties of the criteria? More specifically, the answer to the following questions would provide useful information to applied scaling literature. First, do the differences between empirically and factor analytically keyed biodata, regarding initial validity estimates and cross-validities, that have been documented previously (Mitchell & Klimoski, 1982) remain consistent across different developmental sample sizes? Second, pertaining to the relationship between initial validities and cross-validities, how does a biodata key created using a multidimensional item response theoretic (MIRT) technique compare to other more common procedures? Third, how does the MIRT key fare across a number of developmental sample sizes?

Biodata has been used in a performance prediction context for over one hundred years (Owens, 1976; Stokes, 1994). Due to impressive validities for predicting training and job-related success, biodata has been a popular method for collecting and using background information. Though most would acknowledge without hesitation that biodata are quite useful, few would agree on exactly why they were so. Additionally, exactly how biodata fit into modern psychological theory is a subject of much controversy (Mitchell, 1998). This uncertainty extends into the realm of biodata item information use,

which has manifested itself in a myriad of presently available methods for keying biographical data.

There are a number of different approaches to item keying, whose reasons for development and continued use depend mainly on the purpose for which they are used. Empirical methods, by far the oldest based on the literature (Goldsmith, 1922), were developed to maximize the relationship between responses to items and the criterion of interest. Due to the fact that items are included in a scale based solely on their criterial relationship, the heavy use of empirical keying procedures in the early days of biodata research led to the field being characterized as "dustbowl empiricism." To combat this, recent development has focused on the use of procedures that are more amenable to construct development and theory building. These methods include factor analytic and rational keying, and subgrouping procedures for clustering individuals based on profile similarity.

How do these different methods compare from a psychometric perspective? Studies that have compared various methods have tended to fall into two camps. The first, which includes efforts such as that by Devlin, Abrahams, and Edwards (1992), focus on comparing various empirical keying procedures for predictive efficacy and validity coefficient stability. The second, characterized by studies like Mitchell and Klimoski (1982), compare empirical methods to alternative strategies. One of the major weaknesses of the studies falling into either of the two areas (the aforementioned study by Devlin, et al., 1992 is a notable exception) is that rarely is there any comparison of procedures across a variety of different key developmental sample sizes. This lack of

research information leaves a hole in our current sphere of knowledge that needs to be filled.

Another shortcoming of current biodata technology is the conspicuous absence of modern measurement theoretic utilization. Though some recent efforts have employed the use of confirmatory factor analysis or structural equation modeling (Schoenfeldt & Mendoza, 1994), the area of item response theory (IRT) has not seemed to have much of an impact on research in biodata. Tenopyr (1994) stated this situation has probably resulted from stringent assumptions and prohibitive sample size requirements that are a necessary evil in using IRT. One of the assumptions Tenopyr (1994) referred to was that of "inherent unidimensionality," which pertains to the fact that most of the highly used IRT models require items scaled together be representative of a unidimensional construct. The recent development in IRT models (Ackerman, 1994) that permit multidimensional latencies underlying item responses provides promise for a number of areas using sophisticated measurement methodology. The application of IRT methodology to biodata is long overdue, and research exploring this area will provide an important addition to the scaling literature.

This study addressed two areas of importance to biodata research. Following Devlin et al's (1992) suggestion, a comparison of biodata scaling methods across a variety of sample sizes was performed. The other area, applying modern state-of-the-art measurement methodology to biodata applications, was accomplished by using an MIRT-derived scale as one of the keys compared. This study addresses those two issues by comparing different keys for stability and accuracy of prediction. Keys developed using MIRT were compared with those developed with a linear multidimensional technique

(factor analysis) and a traditional empirical keying method. The criteria of interest were success in a training program and on-the-job (OJT) field performance. In the spirit of Devlin, et al. (1992), the key development sample sizes were varied systematically. In addition to examining the effect of developmental sample sizes on the stability and magnitude of the correlations of the scales with the criteria of interest, each key was assessed for its ability to add incremental predictive power over and above cognitive ability. Predictive power was defined as the ability to correctly assign individuals to the pass/no pass conditions.

In line with past research (Fuentes, et al., 1989), it was hypothesized that the empirical key would have higher initial validities than the factorially derived scale, but that upon cross-validation; the difference between the two would be minimal. It was also hypothesized that the magnitude of differences between initial validities will be considerably higher at smaller sample sizes due to greater capitalization on chance (extraneous variance) in the empirically derived key. Due to the absence of studies comparing the results of classically derived indices versus those obtained via IRT, it was difficult to make predictions on the comparison between the MIRT and other scaling methods. Studies using IRT are still quite sparse in industrial/organizational psychology literature (Guion, 1998) and biodata research in particular. This project served to demonstrate the potential benefits of modern test theoretic methods.

CHAPTER 2 - LITERATURE REVIEW

The assertion is often made that individuals are the sum total of their behavior and experience (Allport, 1937). This notion, along with the widely held belief that the best predictor of future behavior is past behavior, is at the core of the keen interest in biographical life history information. Information on life history can be obtained in many ways, including narrative biographies, interviews, cumulative observational records, and biographical data questionnaires. The latter, referred to as biodata, have been a preferred method for gathering life history information in applied psychology for over one hundred years (Stokes, 1994).

Historical Overview

The first known use of the method was the "job application blank" introduced in 1894 at a meeting of the Chicago Underwriters. A series of standard questions were proposed assessing key elements of an individual's life experience used to improve selection of life insurance agents. Examples of the types of questions suggested included marital status, present and past addresses, individual financial status, and previous work experience (Owens, 1976).

From the early part of the twentieth century until World War II, a number of publications dealt with empirical analyses of biodata item responses for sales and other occupations (Goldsmith, 1922; Russell & Cope, 1925). These focused on discrepancies between responses of good and poor performers and subsequent weighting of item responses for generating predictor scores (Stokes, 1994). During World War II many studies (cited in Owens, 1976) conducted by and for the military reported impressive validities of keyed multiple-choice items for predicting a number of organizationally

relevant criteria. Among these were success in training, post-training performance ratings, and attrition. In one study (cited in Owens, 1976), scored biodata were found to be more predictive of ROTC leadership ratings for officers and cadets than any combination of ten tests of aptitude, attitude, or physical ability. The scored biodata form enjoyed increasing popularity during the post-war years in both the military and civilian sects (Cowles & Dailey, 1949; Hadley, 1944; Johnson, 1944; Keating, Paterson, & Stone, 1950; Levine & Zachert, 1951; Lockman, 1954; Mock, 1947; Mosel & Cozan, 1952; National Research Council, 1946).

One of biodata's strengths (and weaknesses) has been its ability to predict future performance. Until the 1960's, the primary focus was on the construction of items for maximizing criteria prediction. Numerous articles and reports described methods of keying responses to particular sets of questions used in conjunction with selection systems. By 1935, Long and Sandiford (1935) were able to cite over 20 different methods for empirically keying item responses. Capitalizing on this strength was the popular weighted application blank (England, 1971). Though a great deal of work in practice and research dealt with biodata's virtues as a predictor, little if any progress was made in the area of theory development. Whether it was explicitly stated or not, the implication was that it didn't really matter why biodata worked, the important thing was that it did.

Another current of thought was running through the post-World War II psychological community, however, that realized the importance of developing a theoretical footing for future biodata research and use. Perhaps due to the relaxed mood that existed in the United States, as a result of the enviable world position that was

occupied, during the immediate post-war period; more time was made available for scientific inquiry that was not directed toward national crisis. This was also a period of time that proved to be a major crossroads for psychology as a whole. Psychologists had proven their worth during World War II, as they had during the "Great War," but due to numerous fissures that had developed between those practicing psychology and those teaching psychology the field as a whole was trying to "refind itself." The most outwardly noticeable sign of this was the reorganization of the American Psychological Association, for the purpose of remaining "the organization" representing the psychologists in this country.

During this period another closely related area to biodata was seeing large gains in the area of theory development. Influenced by the state of learning theory at the time, and using factor analysis methods, many of the building blocks that characterize our current conception of the field of personality were laid (Pervin, 1990). In addition, the "cognitive revolution" marked the beginnings of what could be characterized as the waking of a sleeping giant. Amidst this backdrop, it was no wonder that those who had admired biodata's more utilitarian qualities in practice were moving in the direction of a conceptual foundation for measures of life history.

Paving the way were a number of theoretical works extolling the necessity of using scored life history questionnaires (Owens & Henry, 1966). In his now classic treatise on "the two disciplines of scientific psychology," Cronbach (1957) proposed the schism between experimental and correlational psychology could be mended via the use of biodata and other psychometric information. According to Cronbach (1957), there was a need for historic information in the entire field of measurement for increasing

understanding and permitting enlightened inferences of causation (Owens & Henry, 1966). Tyler (1959) stressed a need for studying human choice behavior in conceptualizing individuality. She pointed out the efficacy of biographical information for inferring patterns of differential choice behavior across the life span; thereby increasing individual predictability and understanding. Addressing the issue of improving the prediction of criteria, others (Dunnette, 1963a; Ghiselli, 1956), were optimistic about the potential benefits of sub-grouping analysis based on information obtained from biodata.

A number of key events in the development of biodata occurred during the 1960's. Spearheaded by the direction of William A. Owens and associates, major strides were made in the area of categorizing and cataloging scored multiple-choice life history items (Glennon, Albright, & Owens, 1966). As a product of these efforts, an exhaustive list included items tapping areas such as school and work, personal relationships, health, and attitudes, among others. In addition, Owens and Henry (1966) provided one of the earliest overviews of scored autobiographical measures, which included a review of previous efforts up to that point, recommendations on item construction, psychometric properties, and then-current and potential uses. The climax of this period, however, was a conference (Henry, 1966) that brought together leading individuals in the field for the purpose of defining the past, present, and future of biodata as a discipline of inquiry. The conference served to bridge the gap existing between the practical and the burgeoning theoretical foundations, and provided the impetus for development of modern biodata research.

The decades that followed have proven to be very fruitful ones in terms of establishing an understanding of the nature of biodata and providing guidelines for its usage. Wernimont and Campbell (1968) proposed a "consistency model" that took the emphasis in employee selection away from an almost total reliance on tests as predictors. Their model's essence was "the establishment of consistencies between relevant dimensions of job-behavior and pre-employment samples obtained from real or simulated situations." The new procedure placed a considerable emphasis on the use of background data (Wernimont & Campbell, 1968). Concurrently, Owens (1968), using Cronbach's (1957) theoretical "one discipline" network as a point of departure, presented his developmental-integrative model for the first time. This model, which was originally proposed as a way of aiding the integration of experimental and correlational (or individual difference-based) disciplines of scientific psychology, established a framework for using biographical information to "discover" subgroups of individuals displaying differential development. Knowledge of these different patterns would then be used to understand and predict future behavior. Subsequent empirical work (Owens, 1971, 1976; Owens & Schoenfeldt, 1979) served to solidify the potential benefits of the model. On a more practical level, Asher (1972) provided some guidelines for defining what biodata should and should not be, and Thayer (1977) described the evolution of a then 55-year old biodata instrument that had been used successfully in the life insurance industry.

What is Biodata?

Before proceeding with further discussion, it is important to define biodata and the attributes of items that fall under this rubric. Henry (1966) stated, this task has been difficult due to the large amount of controversy surrounding it. For the purpose of this

research, the working definition of biodata is that it "is a measurement strategy that is deeply rooted in the past behaviors and experiences of the individual" (Nickels, 1994, p.2). Nickels (1994, p.2) specifies by adding that "biodata items require people to describe behaviors and events occurring earlier in their lives." As Nickels (1994) has pointed out, many researchers (Asher, 1972; Mumford & Owens, 1987) attempted to establish guidelines for defining exactly what is and what is not biodata, however, the universality of these efforts has not been accepted.

Mael (1991) provides the most recent attempt to pull together the current streams of research trying to establish a common framework for biodata research, by codifying attributes. Mael's synthesis of the current state of knowledge on what constitutes biodata item attributes was presented in tabular form (Mael, 1991, p. 773). The characterization presented drew on the work of others (Asher, 1972), but included revisions to reflect the state of current research, and sensitivity to aspects affected by legal and social concerns. Mael (1991) defined ten attributes or dimensions that fell into three broad categories: *historical, methodological,* and *legal/moral.* Mael (1991) mentions that an additional category of attributes that has received attention in the past (Owens, 1976) concerned response scale alternatives. Mael (1991) also points out that though this particular area is of great importance, the key points apply to all self-report measurement, and are therefore out of the scope of his discussion.

The *historical* category encompassed that dimension of biodata that many would see as the defining characteristic that separates biodata from other domains (Gunter, Furnham, & Drakeley, 1993; Mael, 1991; Nickels, 1994). Biodata has not been consistently defined with this aspect in mind, especially in earlier times when there was a

tendency among many researchers to label any personal information (i.e., personality) as autobiographical self-report data (Owens, 1976). By limiting biodata to events that have taken place or continue to take place, while excluding items about hypothetical behavioral intentions, the possibility of a respondent fictionalizing himself is speculated to be reduced (Asher, 1972).

The issue of controlling for fallacious self-presentation is, however, more directly addressed via dimensions that fall under the second category (*methodological*) of item attributes (Mael, 1991). In fact, Mael (1991) orders these dimensions in such a way that they form a rough continuum with each attribute setting a higher standard for ensuring self-report accuracy. *Externality* refers to the extent to which behaviors in a particular item could have been witnessed by outside observers. Mael (1991) provided the example of a question soliciting information on whether respondents had been fired from a job as one that would have a high degree of this attribute. An item dealing with individuals' attitudes toward marijuana smoking would not, however, be externally observable (Mael, 1991). A closely related attribute pertains to the *objectivity* of the events described in the item (Nickels, 1994). Whereas asking the number of hours spent preparing for a dissertation would be quite objective, asking about the respondent's feelings during that time would not. Furthermore, a high degree of *first-handedness* (Nickels, 1994) would reduce the possibility of response distortion. Inquiring about an individual's typical attendance at work rather than what significant others (e.g., supervisors or co-workers) would say about the respondent's work attendance would minimize speculation that goes into providing a response.

The fourth attribute in this category, according to Mael (1991), deals with the *discreteness* of the item information. This refers to a single, unique event or a simple count of unique events, as opposed to summary (e.g., average number of hours spent engaging in a particular event). Mael (1991) posited this attribute may be desirable because it only requires memory retrieval, whereas, summary estimations require a greater degree of cognitive tasking, which increases the likelihood of inaccuracy. However, Mael (1991) did not negate the potential usefulness of summary measures, particularly with regard to prediction of "typical" performance.

Finally, the *verifiability*, or extent to which a respondent's answers can be substantiated by outside sources is an important dimension of biodata. Mael (1991) noted low consensus regarding the importance of verifiability as a criterion to be met for biodata. Some researchers (Asher, 1972; Guion, 1965) place a great deal of importance on this attribute, while others (England, 1971; Mumford & Stokes, 1992) take a more relaxed stance. Mael (1991) stated the requirement for item information verifiability might be better defined as "verifiable in principle." Here he noted actual verification of a large number of items might be costly and impractical, which would cut into the benefits of biodata use. The value added might actually lie in the respondent's perception that his answers could be "checked for accuracy" rather than whether or not they were actually subjected to this test.

The final four attributes of biodata items (Mael, 1991) are those that pertain to *legal* and *moral* issues. These characteristics of biodata items are most effected by the contemporary legal and social climate and open biodata up to the most public scrutiny

(Farmer & Witt, 1998). It is here that biodata's survival as an applied instrument for decision-making (i.e., employee selection) may lie.

The first of these, *controllability* refers to the extent to which the information obtained in a particular item is a function of the respondent's direct control. As Mael (1991) noted, this attribute is directly related to the conceptual foundation for delineating between *input variable* and *prior behaviors* item types (Owens & Schoenfeldt, 1979). Controllability as a characteristic of biodata is an area that is subject, as many others, to being at the mercy of a double-edged sword. From a theoretical perspective (Mael, 1991; Mumford & Stokes, 1992; Owens & Schoenfeldt, 1979) the amount of individual control over past events should not be at issue. The things that "one does" will not necessarily affect or shape later behavior any more than the things that "are done to one." Whereas, an individual's choice to participate in a particular activity is essentially a function of a decision that is consciously made, the fact that the individual's parents participated in the same activity may be exerting "indirect" control on the ultimate behavior. Further, each or both of these aspects can serve as future behavior shapers. Though demographic variables (i.e., SES, race, gender) are often times frowned upon (Mumford & Stokes, 1992) as biodata items, they too can serve to shape subsequent behavior and would merit consideration in any theoretical discussion on the effect of past events on future behavior. Even a cursory perusal of the content of many biodata instruments used in practice (England, 1971; Glennon, Albright, & Owens, 1966; Mael, 1991) yields a substantial number of items that are definitely not under the direct control of the respondent.

Mael (1991) pointed out that when noncontrollable items are used in situations where important decisions are at stake (e.g., employment), arguments based on theoretical

reasoning lose out to legal reality. It is well known that such variables as gender and race are definitely "off limits" when considering an individual for employment. Similarly, practitioners often advise that "any" variable dealing with demographic, parental, or childhood information be excluded from a functioning biodata instrument (Mael, 1991). Though perhaps quelling certain ethical concerns, it should be noted (Mael, 1991) that totally eliminating noncontrollable variables can oftentimes lead to undesirable properties. Mael (1991) cites an example where noncontrollable items were excluded from an assessment profile on leadership effectiveness. Due to the fact that a complete assessment of the relevant domains was made untenable, the researchers were forced to include behavioral intention-type items. Ultimately, the decision to limit the controllability factor of items boils down to the intended purpose of the instrument, with special attention given to potential legal concerns.

Highly related to the controllability attribute is that of *equal accessibility*. Quite simply, this refers to the extent to which the events or experiences are equally accessible to all respondents (Nickels, 1994). An example of an accessibility-related item would be to ask about home personal computer usage, when the implication is that those who are socioeconomically challenged would have no access to computers. Strict adherence to an equal accessibility criterion for biodata item inclusion is neither universally accepted practice or theoretically prudent (Mael, 1991). Differing philosophies of the goal of biodata will ultimately determine the appropriateness of items that potentially discriminate based on accessibility. Legally speaking it may be "safer" to avoid such items, whereas, regarding theory development, past accessibility would be an important determinant of future behaviors (Mael, 1991). This issue is not entirely split on practical

vs. theoretical lines, however, as some practitioners (Gandy, Dye, & MacLane, 1994) strongly advocate continued use of items that may present material that will not be accessible to all applicants. Mael (1991) concluded by stating, "...neither noncontrollable items nor nonequally accessible items need be intrinsically unfair or unethical. Moreover, adopting these constraints would so limit the use of external and objective items under some conditions that one would be forced to fall back on more subjective and fakable ones" (Mael, 1991, p.781).

Another area of definition pertains to the *job relatedness* (Mael, 1991), or as Nickels (1994) puts it, the *situational relevance* of the content of a particular item. As mentioned earlier, from a theoretical perspective, any singular event, patterns of events, or unconscious variable that has occurred in an individual's past can serve as a determinant, or at the very least a moderator, of future behaviors. This, however, can lead to problems for both researchers and practitioners. Though the explanation of a relationship between an apparently unrelated predictor and a criterion may be limited by the capabilities of the researcher (Farmer & Witt, 1998), it is imperative that a rational link be established at some level.

From an applied perspective, this potential ambiguity is subject to legal and public scrutiny, via interpretation of the *Uniform Guidelines on Employee Selection Procedures* (EEOC, 1978). Pace and Schoenfeldt (1977) pointed out that although the usual interpretation of job relatedness equates with criterion-related validity, that knowledge of the fact that content validity evidence, as assessed via job analysis, has played a role in court decisions (e.g., *Watson v. Ft Worth Bank & Trust*) (Ledvinka & Scarpello, 1992), practitioners should be cognizant of rational considerations in predictor-criterion links. In

light of potential ramifications, at least in the public sector, Gandy, Dye, and MacLane (1994) recommended that items show a face valid relationship with elements of the job and, more conservatively, demonstrate an item-by-item mapping of predictors to criterion components.

Mael (1991) commented that using a cautious strategy regarding job relatedness would limit items to the *sample* domain (Wernimont & Campbell, 1968). This would make it difficult to predict a criterion for an individual that had never actually engaged in the specified behavior. Though such a plan would ensure a high degree of face validity, the effects of faking come into play as items that are most obviously job relevant are also the most subject to intentional distortion (Mumford & Owens, 1987). Currently most biodata instruments include a range of items that fall into both *sign* and *sample* categories (McDaniel, 1989; Wernimont & Campbell, 1968).

The final attribute in Mael's (1991) taxonomy is that of perceived *invasiveness*. This dimension deals with the extent to which the items in a biodata instrument infringe upon an individual's right to privacy. Again, there appears to be a trade-off between positive and negative, as item types that are perceived as the least invasive are those that are the most hypothetical and subjective. In an effort to establish some guidelines on what exactly constitutes invasiveness in item content, Mael, Connerley, and Morath (1996) found that the four motives that generated the most concern were: a) fear of stigmatization, b) concern about having to recall traumatic events, c) intimacy, and d) religion. Fusilier and Hoyer (1980) found individual perceptions of the amount of control over the uses of information after its disclosure was directly related to feelings of privacy invasion.

Mael (1991) concluded that although many authors have attempted to establish some framework for characterizing biodata (Asher, 1972), none of them have eradicated the confusion that exists among those using biodata (Bliesener, 1996). Although not always the case, it would appear that the only "given" is that biodata items be historical in nature (Mael, 1991; Nickels, 1994). Though some of the attributes seem to focus on the fakability of items, and others are centered on addressing legal concerns, none has been universally accepted as a criterion for limiting what biodata items can and cannot be.

Advantages of Biodata

As mentioned earlier, biodata effectiveness is predicated on the premise that the best way to determine what an individual will do in the future, given no other information, is to know what they have done in the past. This does not imply people will always act in ways that are familiar to them, after all Lewin (1936) recognized behavior is a function of the person *and* their environment. It does capitalize on the rather obvious fact that people are more likely to exhibit behavior that has been previously conditioned. This propensity to elicit particular responses in particular situations, focusing on *typical behavior,* makes biodata an excellent device for forecasting. Biodata shares this characteristic with pre-employment interviews, background checks, and work histories.

Biodata do have some characteristics, however, that offer advantages when compared to the other methods (Mumford & Stokes, 1992). Biodata, unlike the other methods mentioned previously, can be collected in a relatively short period of time and at considerably less cost. Items are presented in a standardized form via paper-and-pencil or computer-based questionnaire. This allows for a potentially large amount of data to be collected on a large number of people, rendering it a far more economical alternative to

lengthier, one-on-one methods. Another advantage biodata has is that standardized formatting allows for responses to be quantified, enhancing interpretability (Mumford & Stokes, 1992). Two other advantages are tied to the objective format of the items. Item content and form, including the substantive "meat" of an item and the way this substance is presented, can be tailored in such a way as to allow the researcher or practitioner a very clearly defined picture of developmental patterns and relationships. Content and form of the stem, along with the additional leverage offered by the prespecified response options, contributes to biodata's utility. Finally, due to the fact that in a given biodata questionnaire all subjects are presented the same items in the same way, the potential for interviewer bias is eliminated. A number of other advantages to biodata exist, and many of these are presented in Owens (1976, pp. 611-612) and Gunter, Furnham, and Drakeley (1993, pp. 39-44).

Relationship of Biodata to Other Domains

Biodata and Personality

Mumford and Stokes (1992) noted biodata items often appear to be variants of the type of questions found in self-report personality inventories. This observation is made all the more palatable when one considers that biodata items are often strong predictors of scores on personality scales (Rawls & Rawls, 1968). Owens (1976) mentioned the results of a study in which factorially derived biodata scales were correlated with a number of personality measures. In addition to impressive relationships between the biodata and personality scales, the multiple *R*'s that resulted (.50 to .60) when personality scales were regressed on biodata scales lended support for the notion that the two possess a high level of shared variance. In another vein, Mumford and Owens (1987) found that biodata

factors resembling the "Big Five" factors of personality (Digman, 1990) emerged. More explicitly, others have categorized biodata, and other measures of life history, as the "method of choice" for evaluating personality in personnel selection (Nunnally, 1959), and assessment (Dailey, 1960).

The aforementioned has led some to assume biodata items are simply another format for measuring personality (Mumford, Snell, & Reiter-Palmon, 1994), or temperament (Buss & Plomin, 1975). This position would certainly be consistent with those (Allport, 1937) who include an individual's experience in their definition of personality. More recently, others (Ashworth, 1989) focused on the distinction between the two being somewhat arbitrary and artificial. If, however, the distinction is made between "hard," verifiable and factual, and "soft," private and unverifiable, biodata (Asher, 1972) a clear delineation exists. In a recent study, Shultz (1996) tested a number of confirmatory factor analytic models of multitrait-multimethod matrices, and found personality and soft biodata items represented one factor, and hard biodata items represented a second. Though unverifiable biodata appear to draw from a common variance source as personality, hard biodata is distinct.

With this in mind, many researchers (Mael, 1991; Mumford & Owens, 1987; Owens, 1976) have tended toward defining biodata in the way in which Asher (1972) defined "hard" biodata, though this is in no way a universal characterization (Mael, 1991; Mael & Schwartz, 1991). When one considers the domains of interest from a measurement perspective, the differences between biodata and personality become evident. Self-report personality items generally solicit information regarding an individual's predisposition or general behavioral tendency toward a particular situational

state. The focus is the individual's disposition, and therefore is limited to personal identity. For example, a typical question that would assess extroversion (Costa & McCrae, 1985) would elicit an individual's extent of agreement with the statement "I really enjoy talking to people."

Biodata items on the other hand, focus on prior behavior and experiences occurring in specific situations (Mumford & Stokes, 1992). Thus, items measuring behaviors and characteristics of individuals other than the respondent might appear as biodata items (Mael, 1991). Also, whereas personality item responses are supposedly influenced only by dispositional factors, biodata items capture aspects of the environment that affect and are affected by the individual. In addition to personal, they are tied to social factors as well (Mael, 1991). Hence, a biodata item that would appear to measure something akin to extroversion might be "How often do you get together with friends?" (Glennon, Albright, & Owens, 1966) with a set of responses indicating the number of times in a given period of time.

Mumford, Snell, and Reiter-Palmon (1994) noted there are, in addition to the specificity and focus in the measures of each, two major points of departure for personality and biodata. The first area concerns the element of choice. Biodata measures often capture behavioral patterns that are explicitly tied to the decisions individuals make when presented with a particular situational stimulus. Personality measures, on the other hand, are not tied to a particular decision or choice, but more to a preference. Second, biodata items often tap into content areas that are probably influenced more by individual knowledge or skills than by personality. In fact, biodata-type items are often used as a

preferred vehicle for accessing job-relevant information (Hough, 1984) necessary to assess knowledge, skills, or abilities (Mumford, Snell, & Reiter-Palmon, 1994).

Biodata, Interests, and Cognitive Abilities

Mumford & Stokes (1992) noted biodata items have demonstrated a certain amount of overlap with vocational interest inventories (Eberhardt & Muchinsky, 1984). By tapping into past occurrences of behavior, especially those that are directly a function of or are related to particular occupations, biodata measures capture key determinants of interests (Mumford & Stokes, 1992). Mumford and Stokes (1992) noted likely relationships with attitudes and values also would exist for biodata.

As Mumford and Stokes (1992) stated, the relationship between biodata and measures of cognitive abilities has received less attention than that for other areas. As they and others (Mitchell, 1994) have pointed out, there is a fundamental difference between cognitive abilities as they are typically defined/measured and the way in which they are captured with biodata. Generally, aptitude or ability measures are constructed in such a way as to elicit maximal performance in a somewhat artificial problem-solving situation. Advocates (Schmidt & Hunter, 1998) of the use of cognitive ability measures emphasize the high validities that consistently result when using them as predictors of future performance. However, others (Mitchell, 1998) are quick to point out that biodata often yield as high if not higher validities as performance predictors than measures of ability. A recent meta-analysis (Bliesener, 1996), based on 116 studies with 165 independent validities, found an estimated validity of .22 for biodata predicting performance after correcting for a number of analyzed artifacts. Biodata are particularly useful in the prediction of typical or "everyday" behavior (Mitchell, 1994). Though

biodata do not typically provide information on the upper bounds for performance, Mumford and Stokes (1992) speculate that they may be tapping into the same variance that measures of practical intelligence (Sternberg, 1985; Wagner & Sternberg, 1985) do. In fact, properly constructed biodata may be the best way to assess the types of intelligence that are actually better predictors of real world outcomes (Gordon, 1997) such as job and life success.

To the extent that common sense (Sternberg, Wagner, Williams, & Horvath, 1995), creativity (Chambers, 1964; Sternberg & Lubart, 1996), or cognitive style (Sternberg & Grigorenko, 1997) would be reflected in developmental events, biodata may offer a potentially useful alternative to more traditional measures. From a research perspective, biodata presents the possibility for investigating the interplay between environmental factors and cognitive functioning (Schooler, 1984), and is particularly suited to longitudinal study (Owens, 1953; Owens, 1966).

Conceptual Framework

Mumford and Owens (1987) noted that biodata measures solicit information regarding specific behavioral responses to particular situations, leads one to the conclusion that whenever an item predicts performance it must represent a correlate or "sign" for later performance. Owens (1976) argued for the study of biodata based on a developmental framework, and pointed out that the key is to find an item or set of items that in some way appear to be connected to the criterion of interest, with the ultimate goal of establishing a developmental linkage. Specifically, the challenge involves locating a set of items that optimally predict a relevant outcome, while providing a meaningful underpinning for empirical relationships.

There are a number of approaches to establishing a pool of items. In one of these, the items may reflect behavioral or developmental patterns that contribute to or appear to be related to differential outcomes, but are not actual representations of the target behaviors. Mumford and Owens (1987) refer to this as an "indirect" approach. Conversely, a "direct" approach involves establishing an itemset that reflects demonstration of the criterion behavior in question. Which approach is used will depend upon the purpose of the instrument. Items developed "indirectly" may be less subject to the effects of response misrepresentation, however, may be difficult to justify in employment situations where demonstration of job relatedness is paramount. Whenever possible, a set of items generated by both approaches would probably be optimal. Following the process of establishing item content domains, the items must be weighted in such a way as to reflect the relative importance of each in accounting for differential patterns of development.

Mumford and others (Mumford & Owens, 1987; Mumford & Stokes, 1992) emphasized that the aforementioned general description of a biodata instrument is dependent on two assumptions. The first is that a biodata scale's ability to predict a particular criterion rests on the extent to which items are considered a comprehensive description of the antecedent causal behaviors and experiences. Another way of stating this would be whether or not item stems and response options capture the essence of all developmental determinants. Second, the establishment of a measurable relationship that the developmental pattern be defined quantitatively. This property also allows for the relative weighting of items as a function of their importance in the developmental schema. Mumford and Owens (1987) stated these two principles account for the

recognized importance of item development and scaling issues in relation to other topics in the biodata literature. In fact, prior to about the mid-1980's the lion's share of the scientifically relevant literature in biodata, outside of validity studies, pertained to these issues (M.D. Mumford, personal communication, February 3, 1999).

Concerning the latter issue, a number of techniques have been used for scaling biodata items (Nickels, 1994). The methods have been used in other areas of questionnaire development where there is no single correct response, including opinionnaires, personality inventories, and attitude surveys (Hornick, James, & Jones, 1977). The methods can be broadly grouped into test-centered and person-centered. Methods that are test-centered include *empirical keying, factorially derived keying,* and *rationally derived keying.* Person-centered methodology focuses on identifying particular recognizable groups of individuals that share certain past experiences and have common profiles. The method is most commonly known as *subgrouping*, and its development is often attributed to Owens (Nickles, 1994).

Although any of these methods can be used, and each has its advantages and disadvantages (Goldberg, 1972; Gunter, Furnham, & Drakeley, 1993; Hase & Goldberg, 1967; Hein & Wesley, 1994; Hogan, 1994; Hornick, James, & Jones, 1977; Hough & Paullin, 1994; Mitchell & Klimoski, 1982; Mumford & Owens, 1987; Mumford & Stokes, 1992; Nickles, 1994; Schoenfeldt & Mendoza, 1994), the strategy used most often has been some form of empirical keying. More specifically, this term denotes any number of different methods for weighting items or response options based on their ability to predict differential patterns in a predefined criterion (Mumford & Owens, 1987; Mumford & Stokes, 1992; Nickles, 1994). Empirically derived, or *externally developed*

as Goldberg (1972) refers to them, are typically created by correlating responses on items with the target criteri(a)on and weighting responses depending on their predictive ability.

The predictive ability of empirical keys is well documented (Hogan, 1994). In fact, whenever the value of biodata is posited, it is to a large extent based on a century of research and practice resting on the foundation of empirical keys. Though in practice some (Mitchell, 1998; Mitchell, 1994; Mitchell & Klimoski, 1982) appear to view the strong statistical relationships of biodata with relevant outcomes as the bottom-line for evaluation, empirically derived measures are not without problems. In fact, an apparent reliance on this type of keyed instrument, in the absence of theoretical justification, has helped to earn the label "dustbowl empiricism."

In light of this, many (Dunnette, 1962; Henry, 1966; Korman, 1968; Mumford & Owens, 1987; Owens, 1976) voiced concerns regarding biodata's place in psychological theory. Since empirically keyed instruments capitalize on a relationship with a specific criterion, their ability to generalize to many phenomena is at the mercy of the criteri(on)a of interest. A broadly defined criterion will lend itself more readily to a generalizable itemset than a narrowly defined one (Mumford & Owens, 1987; Thayer, 1977). From the perspective of the sample(s) used for item development, Schwab and Oliver (1974) pointed out due to the large number of items typically used in biodata validation studies, there is a tremendous propensity to capitalize on chance relationships that may exist. Finally, due to differential factors that may operate in one group of individuals as opposed to another, a strictly empirical approach could be prone to being effected by the relationship of these factors with the criteri(on)a (Pace & Schoenfeldt, 1977; O'Leary, 1973; Mumford & Owens, 1987). Though this last point may be of concern from a

theoretical perspective, it also presents a potentially problematic situation legally and ethically in situations where outcomes that impact people (e.g., employment) are tied to the results of a biodata-scoring key.

Therefore, rather than using blatant empirical methods as the method for keying a set of biodata items, the specification of a well-defined network of antecedent and criterion behaviors is preferred (Nickels, 1994). As Nickels (1994) points out, a number of studies demonstrated items developed with specific hypotheses regarding the relationship of predictors to criteria in mind were far more likely to produce significant relationships than those developed without this theoretical foundation. Mumford and Owens (1987) pointed out that item pools containing items tapping into behaviors other than those relevant to future performance, and those that fail to take into account between group developmental shifts, will mislead instead of enlighten. Russell (1994) provided an excellent "point-of-departure" for those seeking guidance to developing biodata that are both content- and criterion-valid. By providing examples from the personality, vocational choice, and leadership literatures he provides one avenue for a theoretically sound approach to biodata item generation. In a somewhat different fashion, Fine and Cronshaw (1994), and Gunter, Furnham, and Drakeley (1993) focused on the importance of job analyses methods for establishing critical domains to be measured via the biodata itemset.

A number of individuals (Dunnette, 1962; Henry, 1966; Owens, 1976) recommended theoretically sound procedures be used in biodata development, with some (Fine & Cronshaw, 1994; Mumford & Stokes, 1992; Russell, 1994) providing very explicit guidance on how this might be accomplished. Nickels (1994) pointed out

disappointingly that relatively few published studies have actually appeared that have done so. In an early effort, Himmelstein and Blaskovics (1960), investigated a biodata instrument developed based on systematic analysis of what constituted effective combat performance, focusing on risk-taking tendencies. They found the scale correlated .37 and .41 (both $p \leq .01$) with peer rated leadership and combat effectiveness, respectively. More recently, Russell, Mattson, Devlin, and Atwater (1990), published a study in which they had developed biodata items from the retrospective life-history essays of first-year students at the U.S. Naval Academy. Scales, based on pre-specified criteria, were developed and found to be predictive (validation and cross-validation) of military performance, academic performance, and peer ratings of leadership. In a study cited by Nickels (1994), Schoenfeldt and Mendoza (1988) hypothesized a number of dimensions critical for management performance. Using structural equation modeling, they verified the existence of most of their constructs. Though the aforementioned studies could lead to the conclusion that theory-driven biodata construction is still the exception rather than the rule, the possibility exists that the practice is more widespread than apparent. As pointed out by some (Russell, Mattson, Devlin, & Atwater, 1990), researchers are notorious for failing to provide information on how itempools were developed.

In addition to the fact that the documentation of theory/construct-driven biodata use is sparse, there are also very few well-developed models of autobiographical data. In a sense, most if not all of the defining theories in psychology, especially those explaining developmental issues and individual differences could be used as starting points for establishing an understanding of biodata. Similar to the way in which organizations are viewed as entities that derive their identity from the individuals that constitute such

(Schneider, 1987a; Schneider, 1987b; Schneider & Schneider, 1994), individuals can be viewed as a sum total of their experiences (Allport, 1937). Combine this with the oft-stated principle that behavior in a discrete situation is a function of individual differences the person brings to the situation combined with environmental variables (e.g., constraints, opportunities, etc.); and that such can feedback interactionally to shape the person (Magnusson, 1990), therefore influencing future behaviors; and you have a basic model of how biodata operates as such a strong predictor. Though this explanation provides a simple elegance, the actual application of this conceptual approach to explaining biodata has been slow in coming.

In light of this, it is not surprising that at the present time there is only one comprehensive and well-defined model of biodata. In 1991, Mael attributed this model to Owens, Mumford and their associates (Mael, 1991); however, the foundation for this model was actually laid by Cronbach (1957) in his now famous call to fellow psychologists to integrate experimental and correlational perspectives in research and theory development. From this, along with the then currently popular and well established stream of research using between-group differences as the level of analysis (Cattell & Coulter, 1966, Cattell, Coulter, & Tsujioka, 1966; Cleary, 1966; Ghiselli, 1956, 1960a, 1960b; Toops, 1948, 1959), Owens (1968, 1971, 1976) modified Cronbach's (1957) model, into a developmental-integrative model. Actually formulated as a model for research rather than one of theoretical explanation, the model specifies the clustering, or *subgrouping*, of individuals based on profiles created via autobiographical information. After the creation of subgroups any number of criteria where differential behavior would be expected can be related to subgroup membership. The key here is that

relationships of particular predictors to criteria of interest do not form the basis of group membership. Instead, individuals are assigned to groups, or perhaps more accurately pre-existing groups are discovered, based totally on data provided via biodata (which is more often than not found in the predictor space).

As an aside, a number of research publications, including theses and dissertations (Mumford & Stokes, 1992; Owens, 1976), using homogeneous subgroups as the unit of investigation, found subgroup status was predictive of verbal abilities (Eberhard & Owens, 1975), drug use (Strimbu & Schoenfeldt, 1973), over- and underachievement, Rorschach responses, and vocational interests (Mumford & Stokes, 1992). In addition to ongoing research that supported the predictive ability of the technique from a longitudinal perspective (Davis, 1984; Mumford & Owens, 1984; Owens & Schoenfeldt, 1979), subgrouping also served as a basis for "maximal manpower utilization" (Owens & Jewell, 1969; Schoenfeldt, 1974; Brush & Owens, 1979; Morrison, 1977; Feild & Schoenfeldt, 1975a), and served as an alternative to moderator group analysis (Feild, Lissitz, & Schoenfeldt, 1975; Tesser & Lissitz, 1973; Lissitz & Schoenfeldt, 1974; Novick, 1974; Schoenfeldt & Lissitz, 1974; Novick & Jackson, 1974; Owens, 1978).

With regard to development of the aforementioned theoretical framework, the fact that biodata-developed subgroups were so effective in predicting a number of behavioral outcomes was useful. In addition, it provided a methodological tool for understanding individual differences, and a means for matching people with demands of particular situations (i.e., "the right people in the right job"). Of more interest, however, was a pattern evident from the results of several "key" longitudinal studies. In two parts of an extended study, Feild and Schoenfeldt (1975) and Davis (1984) focused on the transitions

from adolescence to the collegiate years, and from college to early adulthood, respectively. Using a canonical discriminant function analysis, Feild and Schoenfeldt (1975) found adolescent experiences accounted for 33 percent of the variance in collegiate experiences. Similarly, Davis (1984), using the same type of analysis found the adolescent derived subgroups accounted for 17 percent of the variance in experiences likely to occur within ten years of graduating from college. Though the impact of the adolescent-defined subgroups diminished as a function of the amount of time between the life history events and subsequent analyses, the fact remained that subgroup membership served as a predictor of future behaviors.

As compelling as the aforementioned results were, a study by Mumford, Stokes, Owens, and Jackson (1990) provided an even more interesting pattern. They examined how those who had been assigned to subgroups (or "prototypes") via a Biographical Questionnaire assessing adolescent experiences administered upon entering college moved through subgroups formed with information obtained from a questionnaire administered just before exiting college (this survey assessed college experiences). Similarly they administered surveys assessing post-college experiences 2-4 years post-college and 6-8 years post-college. Again prototype subgroups were formed with this information. A series of chi-square analyses revealed individuals assigned to adolescent subgroups tended to enter 2 or 3-college subgroups, and further individuals in the college subgroups tended to enter only 2 or 3 of the post-college subgroups. These results supported the contention that as people move through life, the paths they embark on are to a certain extent shaped by the path they are currently on, and paths they have been on in the past.

To explain the patterns that had been observed across the research, Mumford, Stokes, and Owens (1990) developed a general framework that they coined the ecology model. Simply put, the model assumes the individual to be a purposeful entity who seeks to maximize personal adaptation through learning, cognition, and external behavior over a lifetime (Mumford & Stokes, 1992). Throughout a person's life path a number of different forces help to shape individuality. Whether it be heredity or experiential, the organism's outlook (which takes into account sensation, perception, ad cognition) and associated behaviors will be predisposed contingent upon the environment. This makes the explicit conjecture that each person will seek to maximize environmental and internal rewards and will therefore act in particular ways or choose situations that will aid in this maximization. Since a series of environmental reinforcers and actions by the individual will tend to minimize the internal variability of what is deemed rewarding, the behavior of the organism, as demonstrated by choice of successive environments, will be channeled toward personal fulfillment. Further, choice of future reinforcers is dependent upon the present situation. This individual then develops a certain way of attaining goals that is to a large extent based on the past successes and failures of previous goal attainments. In this way, the individual's behavior patterns are shaped to the point that the old axiom that "the best predictor of future behavior is past behavior" becomes a reality.

Based on the findings of Mumford, Stokes, and Owens (1990), that the dimensions of personal classification that appeared to exhibit the most stability were those that explained ways in which the organism actively interacts with its environment or tries to make sense of its environment, the idea of the individual being "active" in his

or her individuation is a core theme to the model. This finding minimizes the influence of factors that "happen to" the individual or may be subconscious to the individual's perception. Accordingly, some individuals will not totally agree with the ecological framework of Mumford (Mumford & Nickels, 1990; Mumford, Stokes, and Owens, 1990). Mael (1991) falls into this category, and points to the present author's concerns as component reasons for his position. In addition to the negligible attention given to things that "happen to" the individual, coined *input variables* by Owens and Schoenfeldt (1979), and subconscious influences, Mael also points to the import of failures in shaping future behavior; in rebut to Mumford, et al's (1990), emphasis on successes. Mael (1991) proposed using social identity theory, where the individual defines self-concept as an interaction between the personal and social identities, as a possible way of filling in the some of the gaps left by the ecology model. Regardless of the model's shortcomings, it does remain the most completely articulated formulation for explaining biodata in terms of a theoretical foundation.

Biodata Item Characteristics and Development

As pointed out by many (Fine & Cronshaw, 1994; Mumford & Owens, 1937; Mumford & Stokes, 1992; Nickels, 1994; Owens, 1976; Russell, 1994) well thought out development and specification of biodata items is crucial to the measurement and evaluation of the constructs in question. As Brown (1994) elucidated, in addition to performance prediction, biodata item development may also serve the purpose of being the foundation for placement decisions, needs analysis, and theory building and testing. This makes it incumbent on the part of the researcher/user to have a well laid out framework for generating items, and determining how responses will be recorded and

evaluated. These considerations are at the heart of establishing criterion, content, and construct-valid measures of developmental patterns. In addition, the practical and theoretical consequences (Messick, 1989) that result from the use of these measures must be paramount.

Buttressed by these considerations are a number of recommendations for ensuring well-grounded measurement of biodata constructs that are theoretically meaningful, psychometrically sound, and practically useful. It should be noted that these considerations are in no way unique to biodata, but are an essential part of any construct-based measurement, especially that which is explicitly linked to criteria performance. As noted by Mumford and Owens (1987), after determining a set of antecedent behaviors and experiences presumed to provide relevant linkages with a criterion of interest, criterion functioning should be defined precisely. This entails a full analysis and specification of particular levels of performance deemed important to capturing the essence of what a criterion is "all about." This may be accomplished in a number of ways, including obtaining information via a) job analysis (Fine & Cronshaw, 1994), b) substantive literature pertaining to the criterion domain (Schoenfeldt & Mendoza, 1994), and c) life history interview data (Russell, 1994).

Following this phase item stems are derived from the predictor-criterion domains and criterion specifications are developed. As mentioned earlier, Mael (1991) gives a thorough summary of biodata item characteristics including recommendations for item construction and usage. Though somewhat different from more cognitively oriented test items, a number of additional considerations for item construction were provided by Osterlind (1989). Mumford and Owens (1987) pointed out that during this stage, areas

that cannot be measured with biodata (e.g., certain cognitive abilities) should be eliminated from consideration. In addition, the item developer should approach item development from an experimental or "hypothesis testing" frame of reference. Underlying each item specification is the implicit assumption that a linkage exists between the item and some specified later behavior.

A variety of item formats have been used in tests of achievement (Osterlind, 1989) that would not be amenable to items tapping biodata constructs. For example, using a matching, sentence completion, or cloze-procedure format, where a correct response is assumed, would not provide the individual completing biodata items freedom to answer in an honest fashion. On the other hand, multiple-choice, true-false, or short answer types, provided the stems were suitable, would be applicable in a biodata context. Though essay-type items offer a wealth of potential with regard to information that could be gained from biodata, the complexities involved in evaluating them precludes their practical use in most situations (Osterlind, 1989). All of the aforementioned types can be categorized into two basic groups: a) selected-response and b) constructed-response. Selected-response formats are those in which a number of alternatives are presented (in achievement tests, one will be deemed the correct response). The most common example of selected-response includes multiple-choice and true-false items, and it is these that have tended to be favored among practitioners of biodata. In constructed-response items, response alternatives are not provided, therefore requiring the respondent to answer with a word, short statement, or essay. Complexities involved in these items make their use in biodata particularly challenging. Pending future research these formats may help to increase our knowledge of biodata functioning (M.D. Mumford, personal communication,

February 3, 1999). Of vital importance is the issue of item response format matching the developmental hypothesis.

Within the selected-response type of item, a number of different formats exist. Owens (1976) focused on seven of these, and provided examples of each. Of those reviewed, the organizing characteristics defining each item were whether a) the item allowed for multiple responses or only one, b) response options were graded along a continuum or not, and c) items provided an escape option (i.e., "does not apply"). For the purpose of scoring, continuum-type items can be viewed as single entities with multiple levels of the behavior or experience addressed by the item stem. For items that do not present response continua, each option must be viewed as an item unto itself. Explicit binary (e.g., those soliciting a true-false response) items are the simplest example of this. Non-continuum items with single or multiple response options are scored in such a way that each option becomes an item. For instance, an item with five possible options could be scored as five items. Of note is the scoring of escape options. These must be considered in light of the information solicited in the item stem, and the other possible responses available. A continuum-response item with escape option could be viewed as two separate items.

Mumford and Stokes (1992) cited three seminal studies addressing issues functional characteristics of alternative item formats. Lecznar and Dailey (1950) conducted a study in which item responses were either scored as a continuum or as separate items. They found that although both methods yielded comparable initial validities, the continuum scored method showed less shrinkage upon cross-validation. Owens, Glennon, and Albright (1962) evaluated item formats for retest consistency and

found that the highest level of consistency was achieved when a) item stems were simple, direct, and neutral in connotation, b) responses were graduated on a continuum, c) response options provided an escape option whenever necessary (Mumford & Stokes, 1992). Finally, Mumford and Stokes (1992) cite a study from 1990 by Reiter-Palmon, Uhlman, and DeFilippo, in which the authors attempted to evaluate item response continuums, via predictive ability, based on the connotation expressed (i.e., negative-positive, limiting-enhancing). They found that an item's ability to predict particular outcomes was a direct function of the connotation expressed by the response continuum. As Owens (1976) and others have pointed out, the preponderance of evidence suggests that the most appropriate format for recording biodata item responses is the continuum-type, with escape option provided.

In addition to the impact of item formats, a number of studies have focused on the issues of biodata accuracy and psychometric soundness. Regarding accuracy, the assumption is often made (Mitchell, 1998) that due to their self-report nature, biodata measures are to be viewed with skepticism. As Mitchell (1998) and others (Mumford & Owens, 1987; Mumford & Stokes, 1992) have been quick to point out; however, the research evidence for biodata's accuracy is favorable. In studies where biodata responses were compared with objective information (Cascio, 1975; Keating, Paterson, & Stone, 1950; Mosel & Cozan, 1952), and non-objective data from those familiar with the respondent (Mumford & Owens, 1987), the amount of agreement was high. In a study investigating biodata accuracy, Shaffer, Saunders, and Owens (1986) investigated responses to a survey and a five-year follow-up, and found that the more objective the item content, the greater the similarity. In the same study they solicited comparative

information from the respondents' parents and found the same pattern (Shaffer, Saunders, & Owens, 1986). Though Klein and Owens (1965) reported that respondents were able to improve their scores when instructed to "fake good", that the effect of misrepresentation was minimized when clear definition of favorable responding was absent. Related to this, Mumford and Owens (1987) cite research indicating scores on a measure of social desirability are related to the responses to biodata items.

From a psychometric perspective, biodata present a sort of conundrum, as they defy some of the more conventional pieces of wisdom. Though a number of studies demonstrate a high degree of retest reliability, the very multidimensional nature of biodata prohibits their evaluation in terms of internal consistency indices. This, in conjunction with the well-known reputation of high criterion-related validity, often appears a riddle to those operating under the notion that a valid instrument must be a reliable instrument. When one considers that the primary method for keying biodata inventories has traditionally been via an empirically based procedure, the high validities make more sense.

Biodata Keying Procedures

The area that has probably received the greatest amount of "press time" for biodata, next to their validities, has been that of item keying. Item keying pertains to the manner in which the data obtained from biodata items will be dealt with on a quantitative level. This data can be manipulated at the item or item-response level, and is generally reflective of the relative importance of particular levels of data to the practical or theoretical utility of a particular biodata instrument.

Currently there are a number of different methods for scaling biodata. From a conceptual perspective, these procedures can generally be grouped via a framework laid out by Brown (1994). Brown (1994) categorized keying methods into those that were externally based and those based on internal information. Externally based procedures use the information obtained from item-criterion relationships; whereas internally based procedures rely on information that exists within a biodata instrument.

Further, each broad category can be subdivided. External methods (which are widely known as empirical methods) fall into two groups that are based on what each item's (or item response's) relationship is to other items. In additive keying methods, the patterns of item responses are meaningless. The important relationship is the one that exists between the individual item and the criterion. Often times, items may be totally independent of each other. Based on their ability to discriminate between those who willor will not be successful on the criterion, items will be weighted in a way that maximizes the discriminability. Typically, predictor scores are arrived at by some method of linear combination. Configural keying methods use the information that exists via the pattern of responses. Individuals are placed into criterion groups based on their configurational pattern. Brown (1994) refers to the latter as interactive.

Internally based scaling encompasses a broad range of methods that include those requiring theoretical knowledge of predictor-criterion relationships, those requiring no knowledge whatsoever, and those that increase predictability by grouping individuals into prototypes. The first type, known as the *a priori dimension method* or rational, as mentioned, require that the item developer have some idea of the conceptual relationships that exist between predictors and criteria. After item development, items are weighted

via expert judgment, internal consistency, or a combination of both (Brown, 1994). On the other side of the coin are those methods that utilize factor analysis or any other dimension discovery techniques to reveal any meaning in the items. The final group of internally based methods is those that require *subgrouping* of individuals into prototypes. After grouping of individuals, each subgroup can be used in further analyses to examine the differential impact of group membership on some criterion of interest.

There are certainly other ways that these methods can be grouped. For instance, methods of scaling can also be divided into those that are person-centered, using individuals or groups of similar individuals as a unit of analysis and those that are test-centered. Of course, by person-centered we are referring to subgrouping. Test-centered (or perhaps more accurately "measuring instrument-centered") focus on the item responses themselves and their relationship to some criterion of interest or to other items. In Goldberg's (1972) extensive review of these methods, he classifies them as *externally developed, internally developed,* or *intuitively developed.* These designations correspond roughly to empirically keyed, factorially derived, and rational approaches.

Regardless of the higher-level classification scheme used for scaling biodata measures, one thing that is evident in the literature is that three of the approaches are relatively recent developments when one considers the chronology of biodata technology. Factorially derived, rational, and subgrouping approaches all appear to have made their debuts in the literature within the last 40 to 50 years. Though none of these methods is explicitly tied to biodata the fact remains that they have initially been used in biodata as a reaction to some of the pitfalls of the most common of all methods for handling biographical information – empirical keying.

Empirical Keying

Schoenfeldt (1996) has commented that the history of biographical data applied to problems of selection corresponds to the chronology of empirical keying. The first scientific reports (Goldsmith, 1922) dealing with the efficacy of biodata were based on empirically keyed predictors used to differentiate criterion performance groups. Typically, empirically keyed items are weighted and summed to form a linear composite. The item weights are derived based on the relationship of each item with the criterion. The purpose of weighting is to establish maximal discrimination between high and low levels of performance.

In 1965, Guion commented "the empirical keying approach appears to be the most commonly employed scoring method when the primary purpose is to maximize the prediction of an external criterion." Though the status of empirical keying as "the" method to handle biographical data has changed in recent years it still remains the most widely practiced method for keying biodata predictors (P.R. Jeanneret, personal communication - July 3, 1997). In fact, there are those (Mitchell, 1998) who would claim that as a predictor of future performance, empirically keyed biodata "cannot be beat."

As previously mentioned, the basis for empirically keyed, or externally derived biodata, rests on the foundation of maximizing the relationship between items and, in turn, itemsets with some criterion of interest. As Thayer (1977) noted, this puts a great deal of emphasis on not only the development and location of predictors, but a great deal of the burden on the specification and development of the criterion. Specifically, the measured criterion should be representative of the outcomes of interest. To date,

however, Mumford and Owens (1987) point out that few studies have devoted much effort to the criterion development side of the coin.

Hogan (1994), citing the work of others (England, 1971), has laid out in cookbook format the seven general steps to the creation of an empirically keyed biodata instrument. In line with the previous paragraph, the first step should always be criterion development and specification. During the course of criterion creation, stipulations for effective performance are established. Once this is accomplished, the second step in the process can take place - the identification of criterion groups (i.e., high vs. low performers). The next three steps are focused primarily on the predictor domain, and involve the selection, specification, and weighting of items. In addition to the differentiation of performance groups, these three steps have probably comprised the bulk of the scientific literature on biodata technology. Following these steps, which collectively form the backbone of key development, the derived key is cross-validated to control for extraneous variance that may have existed in the development sample. Finally, cutoff scores are established to separate those most likely to be successful, with regard to criterion performance, from those least likely.

To a limited degree, descriptions of the first four steps have been provided earlier in this paper. Not described heretofore, however, is the issue of item weighting, or keying. Hogan (1994) posits that it is this step in the process that lies at the heart of empirical keying. This statement is buttressed by the presence of a virtual plethora of articles, beginning with Goldsmith's 1922 paper, dealing with the differential weighting of item responses, items, or itemsets.

A brief review will begin with Long and Sandiford's (1935) monograph. In this lengthy treatise, the authors describe in some detail 21 different methods for establishing item validity (in addition, mention was made of three others). The methods were compared to each other based on ease of computation, ability to discriminate high and low performers, and relative efficiency. Though more focused in terms of the quantity of methods used, comparative research on a then popular method of "conventional" keying, versus a newer "pattern of response" method was carried out as part of the post-World War II Air Force Classification Program (Lecznar, 1951; Lecznar & Dailey, 1950; Levine & Zachert, 1951). Of importance was the finding that a key developed by assigning weights to each option, for items showing a graded pattern of validity coefficients across options, had a higher validity and less shrinkage than a key where only responses demonstrating statistical significance were weighted.

Since these early efforts, a number of methods for deriving empirical keys have appeared in the literature. Probably the most well known of these is the weighted application blank (WAB). The WAB method operates at the item response level, and involves the utilization of those responses that do the best job of differentiating the highest and lowest performing criterion groups (Brown, 1994). In brief, item response percentages for the two criterion groups are calculated, and the difference between the two percentages is used to assign item response weights with difference values being converted to weights based on Strong's (1926, cited in Hogan, 1994) tables. Responses that fail to differentiate criterion groups are removed from scoring consideration. The WAB method is commonly known as the vertical percent scoring procedure.

Other methods that have appeared in the literature include horizontal response scoring, deviate response, and keying based on correlation or regression. The horizontal response technique, similar to the WAB method, consists of dividing the number of high performers endorsing a response by the non-criterion dependent total number endorsing the same response. The resulting percentage value becomes the item weight (Brown, 1994). The deviate response method (Malloy, 1955; Neidt & Malloy, 1954), referred to by later writers (Brown, 1994; Hogan, 1994) as the "deviant" response method, differs from the percent scoring methods in that criterion groups are defined by utilizing the distances of observed scores from the predictor-criterion regression line. A variant of this procedure, the rare response method, requires weighting responses based on the scarcity of endorsement (Hogan, 1994).

Two closely related techniques that rely on the weighting of whole items or the "patterns of responses" (Lecznar, 1951) are the correlation and regression-based methods. The requirements for keying using these methods are that the possible responses to an item be graded along a continuum. Both methods begin with some kind of a correlation coefficient or matrix of coefficients. In the strict correlational method, items are selected and weighted based on their individual bivariate relationship with the criterion. The item weight is then the actual correlation coefficient, or some derivative, representing the relationship between the item and criterion. In the regression method, or more accurately the regression methods, criterion scores are regressed on the items in an inventory. Due to the oftentimes large number of items used in a biodata inventory, an often-encountered problem is that of ensuring an adequate ratio of predictors to subjects (Hogan, 1994). Commonly, derived regression weights are utilized as the scoring weights.

The aforementioned procedures are known as additive methods (Brown, 1994), whereby the scoring procedures produce a linear weighted combination of predictors. First presented by McQuitty (1957), another type of empirical keying are those referred to as configural approaches. These methods do not place the emphasis for predictability on each item, but focus on the pattern of responses to a set of items. In fact, oftentimes, the items themselves possess no useful information. McQuitty (1957) categorizes these methods in three general categories based on the way information is utilized. Other than scant representation in the literature, these techniques have for the most part gone unutilized and appear mainly as a footnote (Brown, 1994). Some new potential uses for this set of approaches seem more likely given the presence of recent advances in modern classification schemes, including CART (Classification And Regression Trees) (Griffith, Frei, Rechenberg, McDaniel, & Snell, 1996) and neural networks (Brown, 1994).

Many writers (Hogan, 1994; Devlin, Abrahams, & Edwards, 1992; Steinhaus & Waters, 1991) have made mention of the fact that there appears to be a relative dearth of studies comparing empirical keying methods. Though this writer agrees that there have not been enough studies looking at different methods systematically (e.g., varying developmental sample sizes), the comparison of empirical keying methods has a history almost as long as the published history of empirical techniques. Long and Sandiford (1935) compared 21 methods of establishing item validities. They concluded, based on computational ease and the fact that most methods produced comparable results, that the biserial correlation was the method of choice. Evident in many of the early studies was a strong "computational efficiency" component when it came to recommendations. In a

way this seems to have been the deciding factor, especially since many of the studies reported that most methods were "equally valid."

Much of the research that followed can be summed up concisely. Regarding the classical response keying procedures, the lion's share of the research has used the vertical percent (WAB) method. England (1971) reported that the vertical percent method would give more stable weights than other available alternatives. Aamodt & Pierce (1987) in a more recent vein, state, "...unless further research demonstrates otherwise, the vertical percent method should be used." Opponents (Lecznar & Dailey, 1950) point out, however, that even though the vertical percent method of keying might yield higher validities initially, that it tends to suffer from a greater amount of shrinkage upon cross-validation than other methods.

When item responses are coded along a graded continuum, there is some evidence that the correlation/regression procedures will outperform those based on response keying. Lecznar and Dailey (1950) found that in a developmental sample, the initial validities of "pattern of response" keys where not as high as those based on a WAB technique, but that on cross-validation they were superior. Gage (1957) found that "logical keys," in which weights followed a graded pattern, were more valid predictors of student-derived teacher ratings than a key developed empirically (i.e., response weighted). Though not utilizing percent keying methods, Steinhaus and Waters (1991) found that a weighted composite based on regression results was a better predictor of attrition 30 months post-hire than other methods studied.

Neidt and Malloy (1954) first reported using the deviate response keying method, and proposed it as a superior alternative to the more commonly used WAB methods.

They found that a key developed by this method added more to the predictive validity of an existing battery than a key developed empirically (WAB). Webb (1960), in a similar study, concluded that although the deviate technique may produce higher initial validities, the cross-validities suffered from a greater amount of shrinkage. Studies involving a variation of the deviate response technique (Hogan, 1994) have produced mixed results. Telenson, Alexander, and Barrett (1983) report that with the rare response method, prediction was superior to the vertical and horizontal percent methods. Aamodt and Pierce (1987), using five different samples, found that keys developed in the traditional WAB fashion predicted better than those based on rare response.

A recent study by Devlin et al. (1992) compared a number of empirically based methods across numerous sample sizes. They found that the most practically significant differences existed at smaller sizes. Here the vertical percent methods consistently outperformed those based on horizontal percent, mean criterion, and phi coefficient methods. At larger sample sizes ($N \geq 100$), the five methods tended to validate and cross-validate at similar magnitudes. They also examined the performance of a rare response key and found it to be a consistently poor performer.

Unfortunately, as previously mentioned, studies that have varied the developmental sample sizes, or other sample characteristics, like Devlin, et al. (1992) are rare. Brown (1994) recommends that more studies comparing different additive methods be conducted. The same would certainly apply to configural methods, given their almost non-existent coverage in the literature.

Another issue of relevance, which at times has apparently generated a fair amount of debate, is the use of unit weights rather than differential weights. Though conventional

wisdom and practice seemed to favor empirically derived, or rational, weighting, a number of studies (Clark & Gee, 1954; Kelleher, 1972; Kuder, 1957; Lawshe and Schucker, 1959; Nash, 1965; Trattener, 1963) concluded that unit weighting produced keys that were as valid as those developed empirically. Though some more recent studies have supported this conclusion for the weighting of predictors (Aamodt & Kimbrough, 1985; Wainer, 1976) and criteria (Fralicx & Raju, 1982), others have suggested that in fact these findings do not provide conclusive evidence (Aamodt & Pierce, 1987; Rozeboom, 1979). Excellent treatments of the weighting issue are available from theoretical (McDonald, 1968) and applied (Stanley & Wang, 1970; Wang & Stanley, 1970) perspectives.

Alternative Scaling Methods

As mentioned, though empirically derived biodata keys have proven to be an excellent predictor of performance, they suffer from one major drawback – lack of a theoretical underpinning. In fact, due to the traditionally strong ties between biographical information and empirically keyed scales, the reputation of biodata as a psychologically respectable tool has suffered. As Schoenfeldt and Mendoza (1994) point out, the desire to preserve the predictive power of biodata, and enhance its theoretical foundation has led to the development of alternative scaling procedures.

Rational scales. Brown (1994) has classified available methods along an "externally vs. internally developed" dichotomy, with the aforementioned empirical methods forming the initial category and alternative methods comprising the latter. Though at a gross-level this would seem appropriate, the classification schema assumes mutual exclusion of methods. In fact, most scaling efforts use a combination of external

and internal information as points of reference during development. This would certainly apply to those techniques that Brown (1994) labels *a priori dimension*, and others have termed *rational* (Hough & Paullin, 1994).

With a rationally oriented development effort, an assumption is made that the researcher/developer has an understanding of the theoretical linkages between future performance and antecedent behaviors/events. As implied, this requires a thorough knowledge of the criterion of interest, which in the case of a job could be gleaned from a well-done job analysis. Utilizing the expertise of subject matter experts (SMEs), biographical items with hypothesized relationships to the criterion will be developed. Included in this effort are an accounting of the relationship of these items to the criterion of interest and a definition of the relative strength or importance of these linkages. Using a purely *theoretical* approach (Goldberg, 1972), items are classified and weighted based on SME judgment alone; with a composite score being a simple summation of the response data for each individual. There is no preclusion, however, that scale development remains totally "empirical-free."

Using the method of internal consistency or "homogeneous" keying (DuBois, Loevinger, & Gleser, 1952; Loevinger, Gleser, & DuBois, 1953), item clusters are formed by grouping those items that have high intercorrelations with one another. These scales are refined by retaining those items that facilitate the highest level of internal consistency, while at the same time minimizing the correlations between items of different clusters. Loevinger, Gleser, and DuBois (1953) reported that developing scales in this manner contributed to maximization of the discriminating power of the instrument if presented in a multiple-score format. From a predictive standpoint (Matteson, 1978), evidence exists

that such an instrument can have comparable validity to one developed by purely empirical means. If such an approach were coupled with periodic empirical checks of the predictor-criterion relationship, the distinction "internally developed" becomes meaningless. Though research (Kilcullen, White, Mumford, & Mack, 1995) indicates that rationally developed scales have the potential for exhibiting a high level of construct validity, and may be less prone to the effects of social desirability than more conventional means of assessing temperament, Brown (1994) points out that little exploration into the nomological networks for targeted constructs has been documented.

Subgrouping. Another method that Brown (1994) classifies as internally developed, which the present author prefers to call "person-centered" is the method of *subgrouping.* A considerable amount of space was devoted to this method in a previous section of this paper; hence a detailed description will not be included here. Two points of importance, however, need to be addressed. The first, a methodological one, is that subgrouping as it has been applied to biodata is not really a technique for scaling as much as it is a way of defining the unit of analysis. Individuals are generally (Mumford & Stokes, 1992) grouped via a clustering procedure of data that has been transformed from latent variable scores (derived through a principal components or factor analytic technique) into inter-individual distance scores. After assignment to groups, further analysis will take membership into account.

This leads into the second point, which is that by its very nature subgrouping may present some problems due to its emphasis on placing individuals into different boxes that are not based on predicted criterion performance, but on between-individual differences that may have apparently no relevance. From a theoretical perspective, this

may not present problems as such and in fact may facilitate a deeper understanding of differential development. Legally, however, especially from the perspective of the employment decision, it may appear to be a form of subgroup norming. For further information on this topic the reader is referred to other sources (Brown, 1994; Gottfredson, 1994; Sackett & Wilk, 1994).

Dimension discovery techniques (Factor analytic). Among internally developed biodata instruments is another class of methods for scale development that has some advantages over those previously mentioned. The class of techniques known as *dimension discovery* (Brown, 1994) includes those methods that do not rely on the relationship of predictors to criteria (though they do not preclude establishing such), but on the relationship of the biodata items themselves to other items in the inventory. Through the use of factor analytic or clustering procedures, those items that share the greatest proportion of variance in common with one another will exhibit similar loadings on common dimensions (as in factor analysis or principal components) or group together in common bundles of items (as in cluster analysis). Thus, through entirely empirical means, the researcher may gain some theoretical understanding of the underlying dimensions for a particular dataset. Individual scores may be calculated for individuals on every dimension, and hence may be used for predictive purposes. Though at one level, the apparent outcome may resemble that of a rationally designed set of scales, the advantage is that no *a priori* knowledge of hypothetical relationships (between items or items and criteria) need exist. As for these techniques versus subgrouping, these methods are typically less labor intensive and, since individuals are considered without regard to group membership, less prone to legal contention.

Brown (1994) noted that there were few if any published examples utilizing clustering or some of the less well-known dimension discovery methods, but mentions many examples where factor analytic and principal components methods were utilized (Baehr & Williams, 1967; Childs & Klimoski, 1986; Klimoski, 1973; Mitchell & Klimoski, 1982). Before proceeding to a more detailed summary of some of the major findings, a few points regarding the similarities and differences between the two are necessary. Both procedures are utilized as a means of reducing a large number of variables to a fewer number of underlying dimensions (termed "factors" in the former and "components" in the latter). At this point, however, the technical differences outweigh the superficial resemblance.

Schoenfeldt and Mendoza (1994) point out that principal components is generally associated with data reduction more so than factor analysis. Each component, rather than being an unobservable latent abstraction, is an observable linear combination of the original variables. The purpose of components analysis is to account for all of the variance in a set of variables as represented by a matrix of intercorrelations. Components are uncorrelated with each other, and account for successively less variance in the data until there are as many components as there are variables. In practice a far fewer number of components than variables will be used to account for the majority of variance in the data and to aid in theoretical understanding. Harman (1976) points out that since the method is so dependent on the total variance of the original variables, it is most suitable when all variables are measured in the same units (or when they have been transformed to standard form so that sample variance is one). An important caveat to the use of principal components is that while all of the variance in a dataset will be accounted for, some of

this may be unreliable. In essence, unless the researcher is sure that measurement error is minimal, there may be a level amount of uncertainty as to what type of variance (reliable or unreliable) is being analyzed (Brown, 1994).

In factor analysis, more accurately referred to as *common factor analysis*, the goal is to estimate a set of underlying dimensions, or common factors, that account for the variance. Factors can be uncorrelated or correlated with one another, and the key difference is that the emphasis is on the "common" variance in a set of items, and not the "total" variance. This fact precludes the relationship between dimensions and individual variables. Since the components model presupposes the accounting of all variance, a regression of item scores on component scores yields the equation:

$$X_{iv} = w_{v1}F_{1i} + w_{v2}F_{2i} + w_{v3}F_{3i} + \ldots + w_{vf}F_{fi} \qquad (1)$$

where X_{iv} is individual i's score on variable v, w_{vf} is the weight for variable v on factor (component) f, and F_{1i} to F_{fi} are subject i's scores on the f factors (Gorsuch, 1983, p. 21). In the common factor model, however, since the emphasis is on the common item variance, it is not assumed that all of the variance in each item will be accounted for by the derived factors. This allows for the presence of "unique" variance in the score of each item, which is analogous to the error variance present in multiple regression. Here, item regression is represented by the equation:

$$X_{iv} = w_{v1}F_{1i} + w_{v2}F_{2i} + w_{v3}F_{3i} + \ldots + w_{vf}F_{fi} + w_{vu}U_{iv} \qquad (2)$$

where model components are as is in the previous equation, with the addition of w_{vu} representing the weight given variable v's unique variance (as embodied in a factor), and U_{iv} is individual i's unique factor score for variable v.

The reader will most likely note that from a psychometric perspective, the common factor model appears to make more sense in that it allows for the presence of latent unaccounted for variance in the score of each item. In fact, the component model is totally unrealistic unless one assumes that all possible item variance is accounted for. However, since the initial goal of using a dimension discovery technique, at least initially, for a set of biodata items is data reduction, the use of principal components may make more sense. Since a number of items may be removed from further use due to the results of a factoring technique, it is important that those items that account for the greatest amount of variance in the dataset be retained (even if it is unreliable). This suggestion becomes more salient when one considers that because the variance in the common and unique factors is unobserved, there is a certain degree of indeterminacy in the common factor model (Schoenfeldt & Mendoza, 1994). For these two reasons, use of a technique based on the common factor approach may appear less than optimal for data reduction, but more suitable than principal components for making sense of a set of variables that will be retained. As an aside, the two approaches often yield similar results in practice (Schoenfeldt & Mendoza, 1994). Though the two are distinct from one another, with separate assumptions made on the data, they are oftentimes referred to using a common framework, and hence will be simply referred to as *factor analysis*, unless reference to a particular model is made.

Factor analysis has been used extensively in many areas of psychology, most notably intelligence (cognitive abilities) and personality. The former area has an unusually intimate relationship with factor analysis in that the common factor model was initially developed largely to explain the structure of intellect. Notable milestones in the

development of factor analysis that have been linked with the study of intelligence have included the seminal work of Spearman (1904), Thurstone (1938, 1947), and Guilford (1956). In the area of personality, the structural models of Cattell (1943), Eysenck (1944), and Guilford (1975) have relied on factor analysis as the means of establishing theoretical understanding. The currently popular five-factor model (Digman, 1990) of personality is based entirely on factor analytic evidence (Digman & Takemoto-Chock, 1981; McCrae & Costa, 1985; Norman, 1963; Tupes & Christal, 1961) and holds a great deal of promise for the utilization of personality as an explanatory variable in performance (Barrick & Mount, 1991).

The published history of the application of factor analytic methods, or some technique attempting to approximate empirical dimensional assessment, can be traced to the era immediately following the Second World War. A number of reports (Berkeley, 1952; DuBois, Loevinger, & Gleser, 1952; Loevinger, Gleser, & DuBois, 1953; Pickrel, 1953) dealt with methods of developing homogeneous clusters of biographical items that were not dependent on the relationship of the items with some criterion. Though these papers did not employ factor analysis per se, they did present methods whose purpose was to form groups of items that were highly correlated with one another, and demonstrated a low-level of interrelationship with items in other groups. Lecznar, Fructer, and Zachert (1951) employed a factor analysis of the Airman Biographical Inventory and discovered a number of factors that captured significant variance that was not accounted for by other tests in the Airman Classification Battery.

Factor analytic studies using biographical information have served a number of purposes. As in the Lecznar, et al. (1951) study, the purpose of the factor analyses in

some of the early studies was primarily as a means of discovering the dimensional structure of a biographical inventory. Morrison, Owens, Glennon, & Albright (1962), using a sample of 418 petroleum research scientists, factor analyzed 75 biographical items that had been shown to discriminate high and low performers on 3 criteria, along with the 3 performance criteria, and extracted 5 factors that accounted for 23% of the variance in the correlation matrix. Others have used the results of the analyses, namely factor/component scores, to differentiate between occupational groups in a sort of crude profile analysis. Baehr and Williams (1967), using the responses of a heterogeneous sample of 680 workers, factored 150 items to yield 15 first-order factors that accounted for 43% of the variance. Further analyses revealed that the mean factor scores were useful in discriminating ten occupational groups. In a similar vein, Klimoski (1973) was able to demonstrate significant differences in the mean factor scores of 3 distinct engineering occupational groups (n=920) based on the responses to a 129-item inventory. Following up their initial study (Baehr & Williams, 1967), Baehr and Williams (1968) also found that factor-score means were useful in differentiating sales managers from salesmen, and high-performing salesmen from those classified as low performers.

A number of studies have utilized factor analyses in prediction. Baehr and Williams (1968) regressed five separate performance criterion measures on biodata factor-score means and obtained multiple-*R*'s ranging from .27 to .50. Childs and Klimoski (1986) regressed three occupational success criteria composites (job, personal, and career success) on to five obtained biodata factors (social orientation, economic stability, work ethic orientation, educational achievement, and interpersonal confidence) and found that the factors accounted for statistically and practically significant

proportions of criterion variance in all three composites. Mitchell and Klimoski (1982) developed a predictive equation via the regression of a real estate sales profession criterion (licensed vs. unlicensed) on six life history factors that were obtained from data collected at initial career training (cross-validated $R^2 = .13$). Using cut scores derived from the mean response frequencies of predicted criterion scores, they were able to correctly retain 68.6% of successes and eliminate 62.4% of failures. Other studies that have used factors derived from biographical data have included those by Morrison (1977), VanDeventer, Taylor, Collins, and Boone (1983), and Neiner and Owens (1982, 1985).

As mentioned previously, the product of factor analytic research that distinguishes it from empirical keying is a somewhat more enlightened view of the dimensional structure of a measuring instrument. With this mind, Mumford and Owens (1987) compiled a list of studies that had used factor analytic techniques with biographical data and were able to locate 21 that appeared in the literature. Focusing on the item content that comprised the derived factors, they found that 26 recognizably distinct factors accounted for the study results. Though the majority of these factors showed up in less than half of the identified studies, six of them appeared in over half of the published literature with two in particular (Personal Adjustment in 18, and Academic Achievement in 16) appearing in nearly all of the studies. The balance of the six frequent factors is comprised of Intellectual and Cultural Pursuits, Introversion versus Extroversion, Social Leadership, and Maturity (Mumford & Owens, 1987).

Though the aforementioned summarizes what has been published, it is limited in that what is really being represented is not a general framework for biodata

understanding, but a summarization of item content that has been popular for biodata practitioners and researchers. Schoenfeldt and Mendoza (1994) provide an example of this point in their description of research conducted by them in the development of a biographical inventory designed to measure customer service orientation. Based on the results of a job analysis, they designed a 137-item inventory that tapped into 16 different areas of relevance. These were grouped into four broad categories that were described as a) Dealing with People, b) Life Outlook, c) Responsibility and Dependability, and d) Catch-all category that included scales for life satisfaction, need for achievement, and parental influence among others. After administration of the survey to customer service employees (n=867), and recoding of noncontinuous and multiple option items, the responses of 240 items were factor analyzed via a principle component procedure with a promax rotation. Ten factors, which accounted for 19 % of the variance in the dataset, were reflective of the *a priori* dimensions, with seven of the ten being directly analogous (Schoenfeldt & Mendoza, 1994). Schoenfeldt and Mendoza (1994) report that though there was some direct overlap with some of the 26 factors that Mumford and Owens (1987) reported, the connection between most of the others was less obvious. Currently, the only information gleaned is that the derived factorial structure is representative of whatever the researcher purports to measure in a given itemset.

Though factorial structure may only provide a general idea of the constructs that underlie a particular set of items, factorial scales can provide the basis for research on measured biodata construct generality and stability. Mumford and Owens (1987) cited research that compared the dimensional structures of datasets in which males and females were factored both separately and together. It was found that less of the dataset

variability was accounted for when males and females were considered together, and that the derived factors were less interpretable. When analyzed separately, it was also found that there were noticeable differences in the content and nature of the dimensions for the sexes (Mumford & Owens, 1987). Similarly, Mumford and Owens (1987) also mention studies in which ethnic group or nationality differences were considered. No differences in the factorial structures for African-American and white respondents were evident, however, gender differences emerged again as females and males (regardless of ethnicity) differed on derived dimensionality. The former finding is consistent with more recent research (Collins & Gleaves, 1998) that found negligible differences in the five-factor model of personality across ethnic groups (African-American and white). Research pertaining to possible differences based on nationality found that of 10 dimensions that emerged, 9 showed up in all of the groups (all from Western cultures) included. Mumford and Owens (1987) asserted that when similar backgrounds are reported, derived factors will be similar, but that prevalent differences in cultural or economic situation may result in different factor structures. They also noted that evidence for gender differences precluded the use of a common factorial structure for describing biographical information.

Regarding the stability of biodata factors, a number of studies have provided evidence that factors may be stable over time. Owens and Schoenfeldt (1979) investigated the predictive characteristics of orthogonal principal components derived from samples (each n=1,000) of male and female college students across five separate years and found them to be essentially stable enough to utilize in independent samples, with minimal loss of predictive power (Schoenfeldt & Mendoza, 1994). Schoenfeldt and

Mendoza (1994) further point out that it is also significant that in the Owens and Schoenfeldt (1979) study, the authors were also predicting a number of criteria, including academic performance, choice of major, and involvement in extracurricular activities. Mumford and Owens (1984) similarly found that factors were generalizable across three different inventories, administered to seven cohorts over a ten-year period of time (Mumford & Owens, 1987; Schoenfeldt & Mendoza, 1994).

In addition to factor generality, research exists that suggests factors may be stable over time. Eberhardt and Muchinsky (1982a) administered a shortened version of the biodata inventory that Owens and Schoenfeldt (1979) had administered, to a sample of college freshmen from a midwestern university ten years after the latter's' original samples. Using orthogonal principal components analysis they were able to identify a similar factor pattern to that of Owens and Schoenfeldt (1979) for males; however, the pattern they obtained for females was markedly different. Notably, factors that had appeared in the Owens and Schoenfeldt (1979) study that were not reproduced included religious activity, sibling friction, and independence/dominance (Eberhardt & Muchinsky, 1982). Eberhardt and Muchinsky (1982a) attributed this apparent change in factorial structure to the changing role of women in society between the two studies. Lautenschlager and Shaffer (1987) reanalyzed the data from both studies and determined that the differences in the factor structures for women was more a function of using the shortened version of Owens and Schoenfeldt's (1979) inventory than any societal role shifts (Schoenfeldt & Mendoza, 1994). When Lautenschlager and Shaffer (1987) used the shortened version to analyze both datasets, similar factor structures for both men and women emerged. Recently, Reiter-Palmon (1996) attempted to replicate the factor

structures for males and females from the original Owens and Schoenfeldt (1979) study on an independent sample of individuals taken 25 years later. Reiter-Palmon (1996) found high stability in the number and content of the factors for both genders, with coefficients of congruence ranging from .72 to .91 for males (13 factors) and .78 to .91 for females (11 factors). She attributed differences that existed to smaller sample size and a more diverse sample in the later group, and to changing family patterns between the two groups.

Studies that have looked at factor stability from a more longitudinal perspective have produced mixed results. Neiner and Owens (1982) administered two similar biodata inventories to the same group of individuals with a seven-year separation (the itemsets differed only in that the first assessed adolescent life experiences and the second early post college experiences). Using canonical correlation to assess the similarity of factors between the two periods, they obtained values ranging from .56 to .64 lending support to dimensional stability even given that the two inventories differed as a function of the different life experiences during the two developmental periods. On the other hand, Mumford and Owens (1987) and Mumford and Stokes (1992) both point to research using data from inventories administered to two different age groups that produced different factor structures. The bottom line is that when the researcher expects that markedly different patterns of life experiences may exist as a function of differential development, different dimensional structures may be expected (Mumford & Stokes, 1992).

Comparison of alternative and empirical strategies. As mentioned previously, a number of studies have compared varieties of empirical keying for predictive ability,

discriminatory power, and efficiency. Though empirical keying has remained the most popular method for scaling biographical information (Mitchell, 1998), a number of studies have compared these empirical keys with some of the alternative approaches discussed previously. Empirical keying suffers from a distinct capitalization on predictor-criterion relationships, and as a result does not lend itself well to the acquisition of theoretical understanding. The other three methods (rational, subgrouping, and factor analytic) systematically attack this shortcoming in different ways. Of the three, subgrouping may offer the greatest potential for predicting a myriad of outcomes and simultaneously furthering our understanding of the dynamic nature of the life experience-developmental relationship (M.D. Mumford, personal communication, October 18, 1999). Unfortunately, other than the research programs of Owens, Schoenfeldt, and Mumford, very little has been done in the way of comparing subgroups with the other keying strategies. However, both rational and factor analytic comparisons have received some attention.

Berkeley (1952) and Pickrel (1953) both defended dissertations that compared homogeneous (rational) and heterogeneous (empirical) keys. Though in initial validation the empirical keys outperformed rational keys for predicting criteria, it was found that the homogeneous keys tended to be more efficient in terms of requiring composites of fewer items to attain validities comparable to heterogeneous keys. Furthermore, the homogeneous keys suffered from a less degree of shrinkage upon cross-validation, and they provided interpretable results; whereas, the empirical key did not lend itself to interpretability.

Hase and Goldberg (1967) compared empirically derived, factor analytically derived, and two variants (one developed via formal theory, the other intuitively) of rationally derived personality keys for predicting 13 diverse criteria in a sample of 200 female college freshmen. They found that all 4 methods produced comparable cross-validities (means for the 4 methods across the 13 criteria ranged from .26 to .27) and that all predicted significantly better than chance. In an extensive follow-up study, Goldberg (1972) investigated five methods of developing keys (the four from the previous study, plus a variant of the factor analytic approach – multiple scalogram analysis) in an attempt to isolate three major sources of variance in a personality inventory. Comparing a) strategy of scale construction, b) number of predictors in predictor function, and c) types of predictor functions utilized, he found no significant differences in the cross-validities of any of the five methods, but did find that the rational and factorial approaches yielded more parsimonious scales. In addition, the factor analytic and rational scales demonstrated a higher degree of fidelity in criteria predicted than the empirical.

Hornick, James, and Jones (1977) investigated the ability of two scales (rational/factorial and empirical) developed from an organizational climate questionnaire for performance criterion prediction. They found that the cross-validity for the rational/factorial key was not significantly lower than an empirically developed key. Furthermore, the rational key was more economical in its use of items. Mitchell and Klimoski (1982) compared empirically and rationally/factorially developed keys in their ability to predict the criterion of real estate sales license attainment. Though the empirical key demonstrated more favorable results upon cross-validation (and accounted

for 8.2 % more of the criterion variance), the rationally developed factorial scales showed a lower degree of shrinkage.

Fuentes, Sawyer, and Greener (1989) compared empirical, factorial, and rational keys for performance prediction in a sample of aircraft pilots. They found that the factorially developed key ($r = .27$) did not differ significantly ($z = .51$, $p < .60$) from that of the empirical key ($r = .32$) and demonstrated less shrinkage. Allworth (1997) found that a rational key demonstrated comparable validity and shrinkage to an empirical key. Hough and Paullin (1994) found very little difference between three (rational, factorial, and empirical) methods in predicting criteria. The differences that did exist tended to favor rational and factorial scales when it came to predicting "predictable" criteria, and empirical scales for predicting "unpredictable." Though Hough and Paullin (1994) appeared to favor rationally developed scales, they admitted that in the absence of well-defined theory, rational scales would not be possible.

This final statement is important because all too often theory is not present in a complete form when dealing with biographical information (M.D. Mumford, personal communication, October 18, 1999). Though a number of recent studies have developed extensive theoretical linkages using biographical information (Kuhnert & Russell, 1990; Mumford, Costanza, Connelly, & Johnson, 1996; Mumford, O'Connor, Clifton, Connelly, & Zaccaro, 1993; Mumford, Uhlman, & Kilcullen, 1992), the vast majority of applications are only partially theoretical if at all (Mitchell, 1998). Mitchell and Klimoski (1982) provide an example where a proposed theoretical foundation was used to develop biodata items (based on a job analysis). By the time the keys were developed, the study focused on a comparison of those that were empirically and factorially developed. Due to

the requirement of strong *a priori* theory required for rationally developed scales, their use is often prohibitive and costly. Subgrouping can be rather labor intensive; however, its practical drawbacks are tied more to potential legal controversy than to methodological considerations. From a strictly practical perspective, empirically keyed and factor analytic scales may be the most widely used (T.W. Mitchell, personal communication, April 1, 1998) in psychology. In situations where no known underlying constructs or predictor-criterion relationships are known, these methods provide rather theory-free access to understanding in a quantifiable manner.

Methodologically, empirical and factor analytic techniques are based to a large extent on classical test theoretic technology. Of significant note is the absence from the biodata literature, or most of the psychological literature outside of education, of measurement models based on latent trait theory. Though it has formally existed in some form for the last fifty years (Lord, 1952), its development and use did not reach fruition until the 1970's. Latent trait theory, more commonly known as item response theory (IRT), is based on the premise that the probability of correctly responding to an item or of attaining a particular response level (in the case of multi-category Likert style items), is based on an individual's possessed level of an directly unmeasurable latent ability or trait. Theoretically an individual's estimated trait level is independent of a particular itempool, and conversely the key characteristics (i.e., item difficulty and item discrimination) of an item are independent of the sample of individuals used in estimation. Tenopyr (1994) noted that the dearth of applications of IRT in psychology was in large part due to prohibitive sample size and itempool homogeneity requirements that fall outside the realm of practicality. Though sample size issues will present problems for many potential

applications, large pools of subjects are not unheard of, especially in large-scale selection or testing situations. The other obstacle, itemset unidimensionality, while once a very real limitation has been overcome to an extent by newer IRT models that not only are robust to, but explicitly model multidimensionality.

Modern Measurement Theory and Multidimensionality

Before proceeding to a discussion of the interplay between currently accepted state-of-the-art measurement theory and multidimensional data analysis, a brief discussion of some of the limitations of classical measurement theory are warranted. All of the aforementioned scaling techniques draw largely from measurement techniques that fall under the rubric of classical test theory. Formalized during the first half of the twentieth century (Lord & Novick, 1968) classical test theory models the observed score (X_O) on a test as a function of an individual's true unobserved score (X_T) and random error variance (E). Commonly this is represented by the equation:

$$X_O = X_T + E \tag{3}$$

Ability is expressed by the true score, which is defined as the expected value of observed performance on the test of interest (Hambleton, Swaminathan, & Rogers, 1991); resulting in an estimate that is expressed as a function of a particular test. Hence, a difficult test will lend itself to underestimating an individual's true score, while an easy one will overestimate this value. At the item level, an item's difficulty level is expressed as the proportion of individuals that correctly answer the item. Of course, this value is sample specific, as a group of high ability individuals are more likely to answer correctly. Likewise, the item's discriminability, or the ability to differentiate high and low ability examinees, is typically represented by an item-test biserial correlation coefficient that is

dependent on the sample of individuals. These test dependant ability and sample specific item parameter estimates are a major drawback when attempting any kind of comparisons between groups or across test administrations, utilizing classical measurement theory.

Additionally, classical test theory depends to a large extent on the concepts of reliability and standard error of measurement. Though in theory, both of these concepts are intended to ensure some semblance of stability and precision, they rest on unrealistic assumptions. Reliability represents the hypothetical correlation that exists between parallel forms of the same test. Though a noble ideal, obtaining parallel tests derived from the universe of all possible tests of a particular construct are near impossible to realize. This results in a plethora of possible coefficients that represent at best a lower bound for estimating reliability, with an indeterminate amount of bias. The standard error of measurement is assumed to be constant across examinees and does not take into account variability of precision across the ability continuum. A final limitation of classical test theory (Hambleton, et al., 1991) is the fact that the unit of measurement becomes the test as opposed to the test item.

Theoretically, these limitations are overcome by the conceptualization of performance in terms of IRT. Directly stated, IRT provides estimates of individual ability and test item characteristics that are test and sample-independent, respectively. Also, reliability is not expressed in terms of parallel measurements, nor is precision limited to an averaged constant index of error. Finally, the unit of measurement becomes the test item, hence the moniker *item response theory* as opposed to *test response theory*. Hambleton, Swaminathan, and Rogers (1991) point out that IRT rests on two basic assumptions. The first of these is that an individual's performance on an item is a

function of an underlying latent trait or ability. The second of these is that the relationship between the latent trait in question and the performance on a particular item can be represented graphically by a function known as an item characteristic curve (ICC). In cumulative models, the ICC is a monotonically increasing function that takes the sigmoid shape familiar to logistic regression modelers. In unfolding models (Coombs, 1964), the ICC resembles a normal distribution. The latter are far less common to those currently familiar with IRT, and have been most useful in attitude and opinion measurement. Figure 1 illustrates a sample ICC for a cumulative model.

- - - - - - - - - - - - - - -

Insert Figure 1 here

- - - - - - - - - - - - - - -

As in classical theory, IRT models individual ability and item characteristics. In IRT, however, these components, or parameters, are estimated in a far more mathematically elegant manner. In addition, another difference between classical and modern theories is that IRT models are by definition falsifiable in that some models may not fit a particular dataset. Similar to structural equation modeling, assessment of model fit is of critical importance in IRT. Standard error of measurement is provided for individual ability estimates in IRT, rather than the single estimate for all individuals provided in the classical model (Hambleton & Swaminathan, 1985). The major distinction between the two, briefly touched on previously, which puts IRT theoretically in an advantageous position relative to classical theory, is the property of parameter invariance. In IRT, ability estimates are independent of the sample of items utilized and

item parameters are non-specific to the sample of examinees. In essence, IRT provides an avenue for near absolute measurement.

There are a number of IRT models available to researchers (van der Linden & Hambleton, 1997); however, three of the most common cumulative models are distinguished primarily by the number of item parameters that are permitted to vary from item to item. The most basic of these is the well-known, one-parameter model, in which the ICC is a simple function of the person, or ability, parameter and one-item parameter, namely item difficulty. The probability of answering an item correctly, or of achieving a particular threshold, is represented by the equation:

$$P(\theta) = \frac{exp^{(\theta - b)}}{1 + exp^{(\theta - b)}} \qquad (4)$$

where: θ = the latent construct of interest (being measured),

$P(\theta)$ = the probability of a specific response given θ, and

b = the difficulty or threshold parameter of the item.

The b parameter conceptually can be thought of as the point on the latent construct scale θ where the probability of a given response is equal to 0.50. In more concise terms, it is the item's location parameter. This model is more commonly referred to as the 'Rasch' model (Hambleton, et al., 1991), and rests on the rather restrictive assumption that the only parameter that will vary in a set of items is the difficulty level.

A more realistic model (R.A.Terry, personal communication, October 18, 1999), albeit a less restrictive one, was originally proposed by Lord (1952), and further

developed by Birnbaum (Lord & Novick, 1968). This model, the two-parameter IRT logistic, is a more general version of the previous and is represented by:

$$P(\theta) = \frac{exp^{Da(\theta - b)}}{1 + exp^{Da(\theta - b)}} \quad (5)$$

where: θ = the latent construct of interest (being measured),

$P(\theta)$ = the probability of a specific response given θ,

D = a constant (most often 1.7),

a = the discrimination parameter of the item, and

b = the difficulty or threshold parameter of the item.

The *a* parameter represents the point on the latent trait scale where the item best discriminates between those of low or high levels of the construct of interest. Graphically, it is represented as the slope of the logistic function (a sigmoid cumulative curve) at the point of inflection.

The final common model is known as the three-parameter model, which adds a *chance probability of success* parameter to the equation and is formulated as:

$$P(\theta) = c + (1 - c) \frac{exp^{Da(\theta - b)}}{1 + exp^{Da(\theta - b)}} \quad (6)$$

where: θ = the latent construct of interest (being measured),

$P(\theta)$ = the probability of a specific response given θ,

D = a constant (most often 1.7),

a = the discrimination parameter of the item,

b = the difficulty or threshold parameter of the item, and

c = the probability that an individual will choose the correct answer, given low ability.

This model is most appropriate in testing situations where a 'correct-incorrect' answer classification is present. Of the three formulations, the two-parameter model would appear to lend itself most to the modeling of items where an explicit "right-wrong" designation is not made (R.A Terry, personal communication, November 8, 1998).

All of the aforementioned common models assume that item responses are categorized in a binary format. This is appropriate for items that are explicitly 'correct-incorrect' or items that indicate 'presence of-absence of' and are consistent with the common binary logistic regression model (Hosmer and Lemeshow, 1989). In practice, however, the possibility of item response alternative possibilities is theoretically limitless, with the majority of items having more than two possible response options. Correspondingly, there are IRT models to deals with the realities of many of these response possibilities. In addition to those mentioned, the recently published *Handbook of Modern Item Response Theory* (van der Linden & Hambleton, 1997) details 28 different models, with allusions to many more that are in development. Some of the better known of these include the *nominal categories model* (Bock, 1972), the *multiple-choice model* (Thissen & Steinberg, 1984), the *rating scale model* (Andrich, 1978), and the *partial credit model* (Masters, 1982).

In line with Owens' (1976) recommendations that the ideal response format for biodata items is the graded continuum, the most suitable IRT model is the Samejima's

(1969, 1972) graded response model. As with the dichotomous response models, the graded response model can be represented as a function of ability and item parameters. However, one other important element in these models is that the response function(s) now represents the probability of choosing one response category over its immediate predecessor given a particular ability level. The item response function, or now category response function, is represented by:

$$P_u(\theta) = P_u^*(\theta) - P_{(u+1)}^*(\theta) \qquad (7)$$

where: $P_u(\theta)$ = the probability of choosing a specific response category given θ,

$P_u^*(\theta)$ = the conditional probability of choosing a specific response category given θ,

$P_{(u+1)}^*(\theta)$ = the conditional probability of choosing the next higher response category given θ.

Graphically, a collection of category response functions can be illustrated by the example in Figure 2. As can be seen, the individual at a given ability level may exhibit multiple response probability option potential.

- - - - - - - - - - - - - - -

Insert Figure 2 here

- - - - - - - - - - - - - - -

Before proceeding to further discussion on IRT analysis, it is necessary to point out that a number of assumptions are understood to be met for a particular model to effectively and accurately represent individual item response patterns. Though the assessment of the degree to which these assumptions are met is not direct, methods for

assessing the goodness-of-fit of a model are indirect means of this goal. An important assumption for many of the more commonly used models is that the items on a particular test or segment of the test are measuring only one ability. This is known as the assumption of *unidimensionality.* In practice, the strict adherence to this assumption can be near impossible, as many factors can influence a set of responses, and therefore more accurately this assumption rests on a dominant factor or ability underlying the item responses. Another important assumption is that of *local independence,* which means that when the abilities underlying test performance are held constant, an individual's responses to any pair of items is statistically independent. This is represented mathematically as:

$$\text{Prob}\,(U_1, U_2, \ldots, U_n|\theta) = P(U_1|\theta)\,P(U_2|\theta)\ldots P(U_n|\theta) \qquad (8)$$

where: U_I = the response of a randomly chosen individual to item i,

$P(U_i|\theta)$ = the probability of a particular response given ability θ.

The intuitive reader will make the connection between unidimensionality and local independence. Stout (1987), recognizing that strict unidimensionality was unrealistic, replaced the similarly viable local independence with *essential independence* that more realistically allows for small correlations between items' responses as opposed to an absolute zero. One final assumption implicit in many IRT models is that performance is not speeded (Hambleton & Swaminathan, 1985). In recent years, however, this assumption has been somewhat relegated to anachronism due to the development of models that explicitly capitalize on test speededness (Roskam, 1997; Verhelst, Verstralen, & Jansen, 1997).

As stated earlier, Tenopyr (1994) noted that there were very few studies, if any, where IRT had been successfully applied to biodata. Given the very important assumption of itemset unidimensionality, this fact is understandable. Though Drasgow and Hulin (1990) point out that, in practice, research has shown 2- and 3- parameter models to be robust to violations of this assumption, the fact remains that "one" dominant dimension is still accounting for the majority of item response variance. Biodata are by their very nature multidimensional (Mitchell, 1996), especially as used in selection situations where the emphasis is placed on getting the maximum amount of explanatory information from a minimum number of items (Mitchell, 1998). This is compounded by the fact that many validational sample sizes are less than optimal for the estimation of item and ability parameters in IRT. Hambleton (1989) points out that, depending on the model used, the number of test items and examinees necessary to achieve stable estimates is variable. For example, Hulin, Lissak, and Drasgow (1982) recommend a minimum test length of 30 items and a sample size of 500 for estimation in the 2-parameter logistic model. As the number of parameters to estimate increase, so do the required minimum number of test items and examinees.

Because of the very real limitations of a large amount of existing biodatasets, the use of IRT has been prohibitive, if not out of the question (C.J. Russell, personal communication, April 2, 1998), for application in the scaling of biographical information. The first of these limitations is less problematic when one considers the use of IRT models that are based on multidimensionality in the data. Until recently, these models had not been useful due to practical limitations in computing power. McDonald's (1967) work on nonlinear factor analysis is considered seminal in this area. Other notable early

efforts include those by Reckase (1972), Mulaik (1972), Sympson (1978), and Whitely (1980). Recently, developments in multidimensional item response theory (MIRT) and associated software have made practical application a much more plausible prospect.

Because of the large number of individuals required for MIRT parameter calibration, at least 2,000 to obtain stable two-dimensional estimates (Ackerman, 1994), practical application is out of the question for most situations. In a large-scale testing or selection program, however, application of this new technology may provide a methodology for increasing the understanding and developing a more psychometrically sound precision-based measurement system. It is conjectured that applying this technique to the scaling of biodata may, in fact, help to address the situation to which Tenopyr (1994) referred. Further, the demonstrated ability to apply current measurement theory to biographical information-type data may provide a step in the direction of moving biodata to more construct-based theory (Russell, 1994).

The Present Study

This study addressed two areas of importance to biodata research. Following Devlin et al's (1992) suggestion, a comparison of biodata scaling methods across a variety of sample sizes was performed. The other area, applying modern state-of-the-art measurement methodology to biodata applications, was accomplished by utilizing an MIRT-derived scale as one of the keys compared. This study addresses those two issues by comparing different keys for stability and accuracy of prediction. Keys developed using MIRT were compared with those developed with a linear multidimensional technique (factor analysis) and a traditional empirical keying method. The criteria of interest were success in a training program and on-the-job (OJT) field performance. In

the spirit of Devlin, et al. (1992), the key development sample sizes were varied systematically. In addition to examining the effect of developmental sample sizes on the stability and magnitude of the correlations of the scales with the criteria of interest, each key was assessed for its ability to add incremental predictive power over and above cognitive ability. Predictive power was defined as the ability to correctly assign individuals to the pass/no pass conditions.

Hypotheses of Interest

A number of hypotheses were generated :

H1: In line with past research (Fuentes, et al., 1989), it was hypothesized that the empirical key would have higher initial validities than the factorially derived scale.

H2: Upon cross-validation; the difference between empirical and factor analytic techniques would be minimal.

H3: It is also hypothesized that the magnitude of differences between initial validities will be considerably higher at smaller sample sizes, due to greater capitalization on chance (extraneous variance) in the empirically derived key.

Due to the absence of studies comparing the results of classically derived indices versus those obtained via IRT, it is difficult to make predictions on the comparison between the MIRT and other scaling methods. Studies utilizing IRT are still quite sparse in industrial/organizational psychology literature (Guion, 1998) and biodata research in particular. From this perspective, the study is largely exploratory in nature.

CHAPTER 3 - METHOD

Sample

Data from candidates for the Federal Aviation Administration (FAA) position of Air Traffic Control Specialist (ATCS) were used for this study. These data were collected from 5,240 prospective ATCSs between 1988 and 1990 and consist of pre-employment screening test scores, biographical information, and FAA Academy performance indicators. The sample represents a portion of those hired to fill the void left by the executive order-mandated termination of over 10,000 ATCSs following the Professional Air Traffic Control Organization (PATCO) strike on August 3, 1981.

For this study, the total sample was 83% male, 93% white, and had an average career entry age of 25.8 years (*sd* = 2.8). The majority of those in the study (55%) had some college experience, 32% had a college degree, 11% had only a high school diploma, and just over 1% claimed having earned an advanced degree prior to entry into the ATCS occupation. Three-quarters of this sample had no prior air traffic control experience before entry into the occupation. All of the candidates were competitively selected, first-time entrants into the FAA Academy.

ATCS job function, requirements, and specifications. Della Rocco, Manning, and Wing (1990) addressed the core function and requirements of the ATCS:

> By definition, a controller is tasked with promoting the safe, orderly and expeditious flow of air traffic. This is accomplished through accurate, effective application of rules and procedures in a real-time, dynamic environment. The current ATCS's job consists of a complex set of tasks that demand a high degree of skill and active application of certain cognitive abilities, such as spatial perception, information processing, reasoning, and decision making (Della Rocco, Manning, & Wing, 1990, p.1).

Harris (1986) identified the critical abilities of the ATCS to be spatial perception, verbal and non-verbal reasoning, and mental manipulation of verbal or numeric concepts. In addition, Broach and Brecht-Clark (1994) noted the importance of short-term memory, movement detection, pattern recognition, and attention allocation. Based on a review of available literature, Harris (1986) concluded neither personality nor temperamental factors were predictive of ATCS performance; however, others (Collins, Nye, & Manning, 1990; Collins, Schroeder, & Nye, 1989; Farmer & Fiedler, 1999; Nye, Schroeder, & Dollar, 1994; Schroeder, Broach, & Young, 1993; VanDeventer, Collins, Manning, Taylor, & Baxter, 1984; VanDeventer, Taylor, Collins, & Boone, 1983) have noted the importance of noncognitive predictors for FAA Academy and field performance.

The actual job tasks of the ATCS are determined by the type of facility that employs them. These facilities can be categorized into three groups: a) Terminal or Tower, b) En Route, or c) Flight Service Station (FSS). "The Terminal ATCS works with aircraft during takeoff and landing, using direct vision, radio communication, or radar to obtain information concerning the position and course of the aircraft, and communicating with pilots via radio" (Sells & Pickrel, 1984, p. 10). En Route controllers monitor the whereabouts of all commercial and some general aviation flights while they are en route to their destination. The En Route controllers are tasked with ensuring proper altitude utilization and legal separation of aircraft. Their primary tool is radar and all communications to pilots instructing them of any necessary ascent or descent in altitude or heading changes are accomplished by radio. Unlike the aforementioned facility types, the FSS facilities' function is not to "control" commercial aircraft pilots but to "advise"

general aviation pilots on weather conditions. In addition to advising FSS controllers assist pilots with accepting and closing flight plans, en route communications for pilots flying visually (visual flight rules or VFR), and originating notices to pilots (Sells & Pickrel, 1984). Since FSS controllers are not responsible for maintaining aircraft separation, and the knowledge, skills, and abilities necessary for successful job performance are different than the terminal or en route controllers, they are selected and trained differently than the other two. The data for this study is only from candidates for the terminal and en route options.

Measures

The Selection of Air Traffic Control Specialists

Before proceeding to a discussion of the actual selection requirements and processes, and a description of the instruments, a brief history of the development of the occupation and the selection for controllers therefore is warranted. As so much of the world as we know it, the development and establishment of the system known as air traffic control is a direct result of the post-World War II economic boom. The commercial impetus for a quicker form of transport for goods and services necessitated the development of an air traffic system to maintain the growing National Airspace System (NAS). Initially, the surplus of ex-military pilots and those with aviation experience was able to meet the manpower needs of the system. Air traffic controllers were originally selected from this group. As time marched on and the demands imposed by the NAS required new able bodies, the original supply of those with military air traffic experience dwindled, it became necessary to consider untrained individuals for placement into the occupation.

Research that had actually begun in the 1940's for military professions (Flanagan, 1947) and further in the mid-1950's for civilians (Hilton & Sells, 1984), coupled with the need to replenish the pool of qualified applicants, led to the establishment of large-scale standardized programs to ensure minimum technical competence. This program, in the form of a centralized school and standardized aptitude screening had been in place since the early 1960's (Hilton & Sells, 1984). The primary selection device implemented in 1964 was a battery composed of commercially acquired and FAA-developed tests that assessed arithmetic reasoning, spatial relations, following oral directions, abstract reasoning, and air traffic problem-solving ability. This test, known as the Civil Service Commission (CSC) selection battery was in place until it was replaced in October 1981 (Collins, Boone, & VanDeventer, 1984; Manning, 1991).

Over the course of some twenty years, minimum test performance standards were varied as a function of previous specialized experience. As research initially suggested that those coming into the civilian occupation with military experience, and subsequently waiving the entry test, had higher field attrition rates, consideration was given to terminating the waiver of entrance exam as a result of experience. The critical discovery here; however, was not that military experience was of little or no value in the civilian air traffic occupation, but that there was a significant effect of age of entry into the occupation, as those with military experience tended to be older, and on subsequent attrition. This led to further research (Collins, Boone, & VanDeventer, 1984; Manning, Kegg, & Collins, 1988; VanDeventer & Baxter, 1984) that established an age effect for air traffic control training and field performance, which concurrently led to and

subsequently supported the Congressionally mandated age cap of 31 years for occupational entry in 1972 (Collins, Boone, & VanDeventer, 1980).

Of particular note is the fact that prior to 1981, the FAA Academy had served almost solely as a training vehicle, with the onus for actual selection being placed on the CSC battery and other pre-employment evaluations, which will be mentioned later. Beginning in 1981, the responsibility of the Academy shifted to training and selection concurrently. With the CSC selection battery, the pre-employment screenout rate was roughly 50% (Collins, Boone, & VanDeventer, 1984). Further, depending on age of entry, field attrition rates ranged from 17% to 42% (Collins, Boone, & VanDeventer, 1984). In order to reduce field attrition and to deal with the large number of occupational applicants following the 1981 strike, the system that was put into place eliminated approximately 90% of those taking the pre-employment test. Subsequently, roughly 30 % to 40 % of those entering the Academy were screened out during training. Estimates on actual field attrition rates have ranged from 5% to 10%, depending on information source and field facility-type (D. Broach, personal communication, March 15, 1993). This has resulted in the maintenance of an active duty ATCS population of between 15,000 and 16,000 controllers.

As stated earlier, the data from this study was collected during the years 1988 to 1990. From 1981 to 1992, the selection system for competitive entry into the occupation was essentially the same, with one exception. From 1981 to 1985, the Academy ran two separate "Screen" programs for terminal and en route specialties, respectively. The training period was approximately 12 weeks. In 1985, this scenario was changed as the

two were combined into one common 9-week Screen, which continued until the discontinuation of the program in 1992, which encompasses the data from this study.

Briefly, the selection program entailed multiple hurdles in which the first was meeting basic eligibility requirements. As outlined in Aul (1998), this included: a) U.S. citizenship, b) 18 to 30 years of age, and c) some combination of undergraduate or graduate-level education, progressively responsible work experience or specialized aviation experience, or both. Following this, applicants were required to take and pass the Office of Personnel Management (OPM) civil service examination, the successor to the CSC, and pass with a score of 70 out of 100. This test will be discussed further in a separate section. In addition the OPM exam, ATCS candidates were interviewed in an attempt to inform the candidate of the rigors of the occupation, and help the interviewer assess the individual's overall suitability for the work. Candidates were also required to meet a thorough medical standards examination and personality assessment, with a focus on targeting emotional instability. Following these procedures and prior to formal employment, the candidate was required to pass a background check that would serve to establish suitability for federal employment and to determine whether or not the individual would receive a security clearance (Aul, 1998). Pending successful completion of the aforementioned, the candidate would be hired as a federal employee and sent to the FAA Academy for training. As an aside, some candidates were also hired through the less well-known noncompetitive programs (Aul, 1998), however, as noted earlier, the majority were hired via the just-described competitive process and hence make up the subject sample pool of this study.

Air Traffic Control Specialist Aptitude Test

As previously mentioned, the written aptitude battery that was put into operation in 1981 was developed to replace the more traditional style CSC that was used between 1964 and 1981. This new battery, hence known as the OPM, was used as the initial occupational qualifying exam from 1981 until 1992. During that period, the test was taken over 400,000 times, with only 25,277 applicants being selected to attend the FAA Academy (Broach, 1998). The battery itself consisted of three parts: a) the Multiplex Controller Aptitude Test (MCAT), b) the Abstract Reasoning Test (ABSR), which had been a part of the CSC, and c) the Occupational Knowledge Test (OKT).

The MCAT was a timed (65 minutes), 110-item, paper-and-pencil test that had been developed to replace the written aptitude instruments used in the CSC. The test tapped into skills that are required to function in the occupation, using a simulated air traffic setting (Manning, 1991). The examinee was provided with a set of air route maps that displayed routes of flight through a sector of airspace (Manning, 1991). A table including other information such as aircraft altitudes, speeds, and planned routes of flight accompanied each map (Figure 3).

- - - - - - - - - - - - - - -

Insert Figure 3 here

- - - - - - - - - - - - - - -

A number of MCAT items required the identification of aircraft that would have conflicts with other aircraft. Other problems involved the computation of time-speed-distance functions, interpretation of tabular and graphical information, and analyzing spatial relations (Manning, 1991). A construct validity study (Harris, 1986) found the test to

have high correlations with cognitive marker tests of Integrative Processes, General Reasoning, Spatial Orientation, Logical Reasoning, and Spatial Scanning (Manning, 1991). Two dominant dimensions underlying performance on the test, perceptual field control and verbal/nonverbal reasoning, were revealed through a factor analysis (Manning, 1991).

The ABSR was the only test retained from the CSC battery (Manning, 1991). It was also timed (35 minutes) and presented the examinee with 50 paper-and-pencil items, which assessed the ability to infer relationships between symbols or letters (Manning, 1991). The test items included letter series and figure classification.

- - - - - - - - - - - - - - -

Insert Figure 4 here

- - - - - - - - - - - - - - -

The OKT was 80 items, timed (50 minutes), and tapped into ATCS job knowledge. The test focused on seven job knowledge areas that were generally relevant to aviation, and particularly air traffic phraseology and procedures (Broach, 1998). Prior to 1981, examinees were awarded extra points for claiming job-related experience (Manning, 1991). Previous measures had been self-reported descriptions of aviation and air traffic experience. The OKT was developed to provide a more objective and reliable picture, and was found to be more predictive of performance in ATCS training than the self-reports (Broach, 1998; Dailey & Pickrel, 1984). The OKT test score was not used to compute an applicant's test score for occupational qualification, but provided additional points for those who already qualified (Broach, 1998; Manning, 1991).

The scoring of the MCAT involved a simple summing of the number of items that the examinee got correct. For the ABSR, the score was obtained via: number correct – (.25) number wrong. The score of the MCAT was weighted 2, and then this weighted sum was added to the ABSR score to form a linear composite. This score was transformed via an OPM transmutation conversion to yield a transmuted composite (TMC) score with a mean of 70 and an upper limit of 100 (Broach, 1998; Manning, 1991; Young, Broach, & Farmer, 1996). The TMC score was used to determine employment eligibility. If the TMC score was 70 or above (75.1 for those without previous experience), the individual qualified for entry into the position at the entry-level grade. Available data from an applicant sample of OPM test batteries taken between 1985 and 1992 (N=170,578) indicate a mean TMC score of 73.30 with a standard deviation of 14.37 (with scores ranging from 19.53 to 100).

At this point, points that accrued due to performance on the OKT (ranging from 0, 5, 10, or 15) were added to the TMC, as well as any veteran's preference (VET) points (0, 5, or 10). This resulted in the final civil service rating (RAT), which in turn formed the basis of an individual's ranking, referral, and selection to attend the FAA Academy (Broach, 1998; Manning, 1991). The candidate's ranking was done at the regional level (there are nine operational regions in the national airspace system), with referral and selection based on regional staffing needs. Due to the high number of individuals applying, testing, and qualifying for the occupation post-strike, the regions were able to select a sufficient number of able bodies with scores generally around 90. Hence, the mean TMC for those entering the Academy (N = 14,392) between 1985 and 1992 is 91.08

with a standard deviation of 5.43 (score range = 70 to 100) (Broach, Farmer, & Young, 1999).

Broach (1998) described the psychometric characteristics of the battery. The test-retest reliability for the MCAT was .60 in a sample of 617 newly hired controllers (Rock, Dailey, Ozur, Boone, & Pickrel, 1981). On the same sample, parallel forms reliability coefficients ranged from .42 to .89 for various combinations of items (Broach, 1998; Rock, et al., 1981). Broach (1998) cited a study (Lilienthal & Pettyjohn, 1981) in which the internal consistencies and item difficulties for ten versions of the MCAT were reported. Cronbach's alphas ranged from .63 to .92, with 7 of the 10 alphas exceeding values of .80 (Broach, 1998). Unfortunately, no comparable data for the ABSR were available (Broach, 1998). As Broach (1998) points out, the available data suggests that the OPM may have been subject to practice effects. VanDeventer (1984) found that increases in OPM test performance due to multiple attempts, were not associated with corresponding increases in Academy scores. Manning (1991) points out that this finding led to an OPM-mandated limit in October, 1985, of one test administration per applicant every 18 months, for those with passing scores.

Broach (1998) also reported the results of validation studies using the OPM test as a predictor of FAA Academy course performance and of subsequent field facility job performance. The study was retrospective in nature and utilized predictor and criteria scores for the 15,875 controllers who survived the Academy and were placed into field training between 1981 and 1992 (Broach, 1998). Validation groups were divided into four samples depending on which version of the Academy training program (1981-1985 or 1985-1992) and which field facility option (terminal or en route) they entered post-

Academy. Results showed that the OPM test rating was a valid predictor of Academy performance in all groups, with uncorrected correlations being statistically significant and ranging from .178 to .222 (Broach, 1998). When corrected for direct and indirect range restriction (Ghiselli, Campbell, & Zedeck, 1981), these correlations ranged from .458 to .502 (Broach, 1998). For field performance, however, the OPM rating to field performance status uncorrected correlations ranged from .014 to .047. Corrections for range restriction resulted in values ranging from .095 to .256 (Broach, 1998). When the corrected matrices were used in a regression analysis of OPM rating as a predictor of FAA Academy and field training status, the results, as expected, yield significant standardized regression weights for the regression of Academy performance on OPM rating for all validation groups. For the regression of field training status onto OPM rating, however, only one of the regression weights (for those trained in the En Route Academy training program who later went on to field training in an en route facility) was significant. Two recent investigations (Broach, Farmer, & Young, 1999; Young, Broach, & Farmer, 1996) utilizing a hierarchical regression procedure (Lautenschlager & Mendoza, 1986) and the Cleary (1968) definition of bias, found significant slope and intercept differences in the regression lines for males vs. females, and whites vs. African-Americans.

FAA Academy Air Traffic Control Specialist "Screen"

In 1976, the FAA implemented a pass/fail training program designed to locate and screen out of the occupation, those persons lacking sufficient potential to become fully certified ATCS's early in their careers (Della Rocco, 1998). This program came into being based on the accumulated knowledge that attrition in field training, post-Academy

was high (43 % failures for en route trainees and 38 % for terminal trainees in 1974) and needed to be reduced. Though the attrition rates for trainees was lower when the CSC was used as the soul selection vehicle than it had been when previous experience alone was the selection criterion, the Congressional Committee on Government Operations concluded in 1975 that the FAA's then-current screening was inadequate (Manning, 1991). At the time, field attrition generally occurred about two to three years into training (Manning, 1991).

It was recommended by the committee that the FAA develop and validate a standardized, centralized program that would serve to further screen individuals following the initial screenouts incurred as a result of the CSC (which was replaced by the OPM in 1981). The object was to reduce training costs, particularly for those who eventually would be unsuccessful anyway. As stated earlier, approximately 40 % of those entering the Academy "Screen" program failed or left for some other reason prior to the point where they would have been assigned to a field facility (Aul, 1998). The program was based on non-radar procedures and from 1976 to 1985 was run as two separate programs for terminal and en route options, respectively. In 1985, the programs were consolidated into a single Screen program, with subsequent field facility assignment being made post-Academy (Manning, 1991). The placement decision was a function of the score obtained in the Screen (Manning, 1991).

The Screen was designed to assess the aptitude of those individuals having no prior knowledge of the ATCS occupation by having them learn general aviation and air traffic control information, and non-radar-based rules and principles with applications. The first segment was strictly academic and covered topics such as principles of flight,

meteorology, the role of air traffic, and the NAS (Della Rocco, 1998). The non-radar segment included in-class academic work, laboratory simulation problems, and a Controller Skills Test (CST). This segment of the Screen was the one that was designed to fulfill the objectives of the congressionally mandated objectives of screening out those with low aptitude.

The non-radar academic portion of this training involved instruction in the rules and principles of non-radar air traffic control. Performance was assessed via multiple-choice testing and the score from this segment constituted 20% of the final non-radar phase score at the time of the Screen's discontinuation. As part of the laboratory training, students applied the rules they'd learned in the classroom to simulated air traffic scenarios. Performance on standardized, timed scenarios, where one student plays the role of pilot and the other "controller", is evaluated by instructors, who are former ATCS's who have been trained as observer/raters. The student's performance rating is comprised of a technical assessment based on the number and types of errors made, and a normative evaluation of the student's performance in relation to others that the instructor has encountered and rated. These scenarios incrementally escalated in complexity and the student's performance on the best five of six graded laboratory problems were combined and averaged to yield the final laboratory performance rating (Manning, 1991), which accounted for 60 % of the final non-radar phase score. Finally the CST, a timed, multiple-choice, paper-and-pencil examination, was designed to assess the ability to apply non-radar air traffic control rules and procedures. This portion made up the final 20 % of the non-radar phase score and the three weighted components were combined to form a 100-point score distribution, with a minimum score of 70 required to pass. Failure to

pass resulted in removal from the ATCS occupation, while success resulted in field facility assignment (Della Rocco, 1998).

Between 1971 and 1992 the FAA was able to reduce attrition from field training from 41% to less than 8% (Della Rocco, 1998; Manning, 1998). As mentioned, this period saw the implementation of a two-option, second-stage hurdle (the FAA Academy Screen) in 1976, the replacement of the initial first hurdle (CSC) with a more effective pre-employment screen-out (OPM) in 1981, and consolidation of the two-option screen into one Non-radar Screen in 1985. Factored into this is the impact that the 1981 strike had on depletion and repopulation of the nation's active ATCS workforce.

A number of studies have been done to assess the validity of the Screen program for predicting attrition rates, supervisor ratings, and field training status (Della Rocco, 1998). VanDeventer (1981) found that the correlation between FAA Academy Screen composite score and field supervisor's rating for the en route option was .56, when corrected for explicit range restriction. Manning, Della Rocco, and Bryant (1989) found correlations of .46 and .30 (corrected) between Academy Screen composite and instructor ratings for the en route and terminal options, respectively. Della Rocco, Manning, and Wing (1990) found an adjusted correlation of .44 between Non-radar Screen composite and field training status in the en route option (Della Rocco, 1998). Further, Broach and Manning (1994) concluded that the Non-radar Screen composite was a valid predictor of succeeding radar training performance. In a final assessment of the Non-radar Screen (Broach, 1998; Della Rocco, 1998), the corrected correlations between Screen composite and field training status at first facility was .25 (N=3,484) for the en route option and .17

(N=2,505) for the terminal option. Della Rocco (1998) indicated that FAA Academy Screen programs might have led to adverse impact against African Americans.

ATCS Field Training

Upon successful completion of the FAA Academy, those who passed were assigned to a field facility for further training (Manning, 1991,1998). At this point, they worked in an apprenticeship-type situation and were referred to as developmentals. During the period of time covered by the present data, assignment to a facility type upon Academy completion was based on Screen performance. The level of complexity at facilities differs as a function of the type of air traffic control services performed and the number and type of aircraft in operation (Manning, 1991). Developmentals undergo training that focuses on the procedures specific to a given facility type and airspace. Training at en route and terminal facilities is segmented into phases with developmental performance being assessed via pass/fail mode (Manning, 1991). The certification of Full Performance Level (FPL) ATCS is awarded after successful performance of all required training in a timely manner. For the en route option, time to FPL averages around 2.9 years; whereas terminal ATCSs time to complete developmental training and become FPL ranges from 1 to 2.3 years, depending on facility level (Manning, 1991). Manning (1998) provides a more detailed description of post-strike field training programs.

For the purpose of this study, an important clarification regarding field facility performance needs to be addressed. Unlike the OPM assessment battery and the FAA Academy screen, field performance is not recorded and indexed via a scaled score. The database for field training contains information on whether a particular phase was passed or failed, dates of occurrence, on-the-job hours to completion, grades, and global

normative instructor ratings (Manning, 1991). In addition, not all developmentals that are unsuccessful at a particular facility or in an option are terminated. Switching to another option or facility level prior to FPL certification was not an uncommon occurrence as 14.2 % of en route and 9.4 % or terminal developmentals who had gone through ATCS training between 1981 and 1992 did so. Data on subsequent performance for these individuals is incomplete. Therefore the data in this study regarding field training status is limited to those who made FPL at their first assigned facility (71.6 % en route, 83.8 % terminal) and those who failed and terminated from the occupation while at their first facility (13.7 % en route, 6.8 % terminal). For simplicity, field training performance is then limited to a pass/fail dichotomy.

Biographical Information

During the years when the FAA Academy ran as a combined option screening program (1985 to 1992) and prior to this, a number of measures were used to obtain information on individual differences that might be used in future selection efforts pending research indicating such. This data was collected from new ATCS hires while they were students at the FAA Academy. Information that was collected included aptitude, temperament, personality, and biographical data. Some of the measures were commercially available instruments, while others were noncopyrighted and developed to assess general non-occupation-specific traits, or those that were specific to the job of ATCS. Instruments utilized to gather biographical information were of the latter type in that they were nonproprietary in nature, and represented both broad non-job- specific developmental markers and those that were derived from analyses of the valid predictors of training performance for ATCSs.

Biographical information as a predictor of performance and attrition has been used by the FAA and its predecessor organization the Civil Aeronautics Administration (CAA) for pilots and later air traffic controllers for a number of years (Brokaw, 1959; Cobb & Nelson, 1974; Collins, Boone, & VanDeventer, 1980; Collins, Nye, & Manning, 1990; Johnson, 1944; Kelly & Ewart, 1942; National Research Council, 1946; VanDeventer, Collins, Manning, Taylor, & Baxter, 1984; VanDeventer, Taylor, Collins, & Boone, 1983). Brokaw (1959) found no appreciable relationship between factors such as marital status, previous flying experience, or education for predicting air traffic control school or on-the-job performance (N=130). More recent studies (Collins, Boone, & VanDeventer, 1980; Collins, Manning, & Taylor, 1984; Taylor, VanDeventer, Collins, & Boone, 1983; VanDeventer, 1983a, 1983b), however, found previous ATCS experience to be predictive of training performance, and that educational level was inversely related to future performance. Of particular interest is that self-reported high school math scores have demonstrated a consistently high correlation with FAA Academy performance (Collins, Boone, & VanDeventer, 1980; Collins, Nye, & Manning, 1990; VanDeventer, Collins, Manning, Taylor, & Baxter, 1984; VanDeventer, Taylor, Collins, & Boone, 1983). As a matter of fact, Collins, Nye and Manning (1990) reported high school math scores to be more predictive ($r = .52$) of Academy performance than the TMC portion of the OPM battery ($r = .42$) (N=3,578). Self-expected level of performance relative to other ATCS's was also found to be highly predictive ($r = .36$). As an aside, all of the aforementioned studies have viewed age as a biographical variable. It was mentioned that age at time of occupational entry has contributed significantly to performance and subsequent attrition

(note the age cap imposed by Congressional Order). For the purposes of this study, age will not be considered.

Biographical Questionnaire (BQ). Regarding ATCSs the majority of the aforementioned studies have utilized an instrument known as the Biographical Questionnaire or BQ. This particular instrument was developed by the FAA and first referred to in a study (VanDeventer, 1983a) from data collected between 1977 and 1979. The original instrument as referred to had 60 items and covered high school education, post-high school education, and prior experience. VanDeventer (1983a) cites Owens' biographical questionnaire (Owens & Schoenfeldt, 1979) as the progenitor to the BQ. Reports of this instrument mention that it had 81 items in 1980 (Taylor, VanDeventer, Collins, & Boone, 1983), 96 items in 1981 (Collins, Manning, & Taylor, 1984), 145 items in 1985 (Collins, Nye, & Manning, 1990), and in its final iteration was expanded to 195 items in 1990. This most recent edition included, in addition to the original content areas, a larger number of items assessing various forms of experience, performance expectations, and extracurricular information. No psychometric information was available on the BQ. The most recent version of the BQ is included as Appendix A.

Applicant Background Assessment (ABA). Between 1988 and 1990, 6,097 students from the FAA Academy were administered the Applicant Background Assessment (ABA) in addition to the BQ. The ABA represents an outgrowth of the research that led to the development of the Individual Achievement Record (Gandy, Dye, & MacLane, 1994), also known as the IAR. The IAR was developed with the intent of providing a relatively broad-base, non-job-specific biodata form to be used in selecting entry-level, nonsupervisory employees for over one hundred professional and

administrative occupations across an array of federal agencies (Gandy, Dye, & MacLane, 1994). A detailed explanation of the development, criterion-related validation, content validity, and methodological issues is presented in Gandy, Dye, and MacLane (1994).

As affirmed by MacLane (personal communication, February 22, 2000), the IAR, and the ABA were developed to tap into rather broad past experiences that were under the control of the respondent and would generalize to a large number of occupations. The ABA is a 142-item questionnaire comprised of four major categories of item content: a) High School Academic Experience, b) Undergraduate College Academic Experience, c) Work Experience, and d) Skills. Interspersed through the academic experience sections are items that dealt with interpersonal relations (Appendix B). The ABA was the predecessor to an OPM-developed biodata instrument that dealt more with issues relevant to ATCS performance, including job-specific criterion validity (C.N. MacLane, personal communication, February 17 & 22, 2000); however, the latter was never released for use.

In 1999, two studies (Dean, 1999; Farmer & Fiedler, 1999) utilizing data from the ABA examined aspects of that instrument. Dean (1999), in an effort to assess the incremental validity for predicting FAA Academy success provided by biodata, found that an empirically derived biodata key correlated .37 (.44 corrected) with Academy performance; whereas the TMC component of the OPM correlated .16 (.42 corrected). It was also found that when the biodata key was added to a predictive hierarchical regression equation with Academy performance regressed on TMC, that the multiple *R* increased by .113. When the TMC component was added to a corresponding equation where Academy performance was regressed on the biodata key, the multiple *R* increased by only .071. In a somewhat different preliminary study, Farmer and Fiedler (1999)

found that a linear composite of biodata items subjectively labeled "Leadership Ability" correlated significantly (R = .56) with the Boldness factor from the 16-Personality Factor (16PF) Questionnaire (Cattell, Eber, & Tatsuoka, 1970). It was also found that the addition of the leadership composite significantly increased R by .03 when added to the equation where Academy performance was regressed on TMC.

Predictors and Criteria

For this study the predictors were the cognitive aptitude component (TMC) score of the OPM battery, biodata keys created from items taken from the BQ and ABA, and the FAA Academy Non-radar Screen composite score. The criteria will be pass/fail status at the FAA Academy, the FAA Academy Non-radar Screen composite score, and FPL/fail status at the first assigned field facility.

Procedures

Sample Management

Of the over 14,000 individuals who went through the FAA Academy Non-radar Screen, 11,405 of these were first-time competitively chosen. Of this group, 5,240 had complete data that included OPM aptitude test scores, Academy performance, and biodata questionnaire responses to both the BQ and ABA. After deletion of those individuals who withdrew from training for a variety of reasons, data from 4,568 individuals remained. It is this data that was used for the biodata key development, validation/cross-validation, and classification analysis for the prediction of FAA Academy performance (Non-radar Screen composite score and pass/fail status). For the 52.1 % of those who passed (35 % failed, 12.8% withdrew) the Academy and entered field training (N = 2,732), field performance data were available for 2,731 (En Route = 1,710; Terminal =

1,021). For the field training, 79.8 % made FPL certification and 8.1 % failed at their first facility. After deletion of those individuals who switched facilities, data for 2,000 individuals remained. These data (OPM scores, Academy scores, and biodata) formed the basis of the key development and analyses for the prediction of FAA field performance status (FPL certification vs. failure).

One of the notable aspects of Devlin et al.'s (1992) study was the comparison of the results of different keying methods for stability based on the size of the total sample used for key development and validation. Extending the work of Devlin et al. (1992), the present study utilized a similar strategy and varied the size of the study samples. Devlin et al. (1992) utilized five samples ranging from 75 to 1200, with a 2-to-1 developmental-to-cross-validation sample size ratio. Due to the large number of subjects suggested (Ackerman, 1994) for one of the procedures (MIRT), this study utilized some samples of larger size. Initial sample sizes of 100, 500, 1000, 1500, 2000 and 3000 were drawn randomly from the original Academy only data, and these and the full 4568 were used for validation. For the field performance database, initial samples of 100, 500, 1000, 1500, and the full 2000 were used for validation. Based on the recommendations of Devlin et al. (1992) and a number of others, a 2-to-1 developmental-to-cross-validation ratio was used. Therefore, each of the aforementioned samples was divided into developmental and cross-validation subsamples. Breakdowns for each subsample by pass/fail and developmental/cross-validation sample are presented in Table 1.

- - - - - - - - - - - - - - -

Insert Table 1 here

- - - - - - - - - - - - - - -

Initial Selection of Biodata

Prior to utilizing the biodata items to create the scoring keys, a subset of the items from the BQ and ABA was culled out to eliminate those items that, in the opinion of the researcher, appeared to fall outside of the definition of "biographical information", solicit very specific aviation-related experience, or demonstrate a high-level of item content overlap between the two questionnaires. Of the 145 items that comprised the BQ at the time that the data for this study was collected, 68 were used. The retained items are indicated by an asterisk next to each item in Appendix A. All of the items from the ABA (item n = 142) were used, resulting in a final set of 210 items that were submitted for possible inclusion in the derived scoring keys.

The purpose of each scaling method is to utilize the information from each item quantitatively, as a basis for establishing individual differences. This entails identifying the relationship between each item and an appropriately defined criterion, or between all of the items, or a combination of the two. Based on this information and other data characteristics (e.g., missing data, item invariance, etc.), items are selected for final inclusion in the scoring key. Retained items are weighted and the rules for scoring are established.

Empirical Key (Vertical Percent Method)

For the empirical key, high- and low-criterion groups were created by using the classifications already existing; those who passed versus those who failed. For the Academy performance validation, this results in 52.1 % success versus 35 % failure. As mentioned, the 12.9 % who withdrew were not included. For field training performance, the representative percentages are 79.8 % success to 8.1 % failed, with the 12.1 % switching to another option or facility being dropped.

The vertical percent method functions most efficiently when there is endorsement of all item-response options by both criterion groups (England, 1971). The differences in endorsement percentages are computed for each alternative by subtracting the low-criterion group percentage from the high-criterion percentage (Devlin, et al., 1992; England, 1971). The difference in percentages is weighted via any of a number of different strategies. For this study, the vertical net unit weight strategy was employed (Devlin, et al., 1992; England, 1971). With this strategy, the percentage point differences were transformed into weights utilizing Strong's (1926) tables (in Stead & Shartle, 1940, p. 255). Next, these derived weights were transformed into values with a range of 0 to 2, with magnitude and sign of the original weights both being considered. An example item weighting scheme is presented in Table 2.

- - - - - - - - - - - - - - -

Insert Table 2 here

- - - - - - - - - - - - - - -

These weights were then applied to all of the scores in a particular validation sample to score the BQ and ABA. If the weights for all item responses for a particular item were

equal, the item was dropped from the scale due to the lack of variance associated with said item. Table 3 presents a summary of the number of items remaining in each scale after the item weighting procedures in each sample.

- - - - - - - - - - - - - - -

Insert Table 3 here

- - - - - - - - - - - - - - -

The score for each item remaining is the weighted item response. These weighted responses are summed to create an overall biodata score for each individual.

Factor Analytic Key

Prior to performing any analysis whatsoever, particular biodata items were recoded to make them amenable to factor analytic procedures. Graded item reponses with ascending category identifiers were the ideal. Items where the category descriptors were in descending order were reverse coded. Category labels of "don't know" or something analogous were recoded as missing responses. Some of the items were recoded as more than one item to compensate for non-continuous response scales or response options that constituted a nominal measurement level. An example of an item that exhibited a combination of continuum-based with nominal measurement is:

My class standing in high school put me in the:

A. top 10%
B. top 33%
C. top 50%
D. top 90%
E. did not graduate from high school

This item was recoded into two items, with one utilizing the first four response categories (subsequently reverse coded), with those responding to the last option as missing data.

For the second generated item, responses in the first four categories were recoded as 1, with responses to the last recoded as 0 (for this item, the response would be the result of a question soliciting whether or not the respondent graduated from high school). An example of a question that would be recoded as five separate items, due to nominal measurement, would be:

The high school subject in which I received my lowest grades was:
A. science
B. math
C. English
D. history/social sciences
E. physical education

Following this recoding of the original 210 items, 296 items were submitted for factor analytic procedures.

After dividing the study samples into developmental and cross-validation subsamples, item frequencies were run to determine the percentage of item responses per item that had missing data. It was determined that due to the large number of variables going into the multivariate analysis, missing data, especially dispersed randomly throughout each dataset, could be problematic. Rather than apply a data imputation technique, it was decided that items with a particular predetermined percentage of missing values would be deleted from further analysis (J.L. Mendoza, personal communication, July 15, 2000). It was determined that after listwise deletion due to missing data, at least 95% of the respondents should remain for further analysis. Upon investigation of each developmental sample, this target was achieved when variables that had at least 2% missing values were deleted.

Following deletion of variables based upon the missing data criteria, variables were also deleted if 90% or more of the valid data for a particular item fell into one response category (Tabachnik & Fidell, 1996). Table 4 presents the number of items remaining in each dataset after deletion of variables based on the aforementioned criteria, along with the percentage of the total possible number of items across the samples.

- - - - - - - - - - - - - - -

Insert Table 4 here

- - - - - - - - - - - - - - -

Across all samples, the mean (rounded to the nearest whole number) of all itemsets was 133, which represented a mean of 44.9% of the each itemset remaining for further analysis.

The next step involved creating a bivariate Pearson product-moment correlation matrix of the remaining variables. Variables that produced no correlations higher than .30 for any bivariate pair (Tabachnik & Fidell, 1996) were deleted from further analysis. Table 5 presents a summary of the number of items remaining in each dataset after this step. Across all samples, the mean (rounded to the nearest whole number) of all itemsets was 115, which represented a mean of 38.9% of the total itemset remaining for further analysis.

- - - - - - - - - - - - - - -

Insert Table 5 here

- - - - - - - - - - - - - - -

Upon completion of this step, the remaining items were factored utilizing the principal component model. *SPSS 7.0* (Norusis, 1994) was used for the analysis. For

each dataset eigenvalues were computed and plotted. Utilizing the scree test (Cattell, 1966), an initial range of component numbers was determined. This range served as the basis for model testing across the twelve samples. For each sample, except the two smallest (initial sample n =100), a range of from 3 to 8 component solutions were evaluated. For the two smaller samples, this range was extended to 9, to facilitate the fact that determining the "break point" on the obtained scree plots was more difficult. After initial factoring, promax and varimax rotations were performed on each set of components. Using the criterion of interpretability, components were retained for further analysis.

At this stage, one final item deletion was performed on each dataset. Items were deleted if they had no loadings that were at least .55 (Tabachnik and Fidell, 1996, characterize this as a "good" loading) with any of the retained factors. The remaining items were refactored using the aforementioned procedure. A summary of the number of items remaining, the retained factor number, and the percent of variance accounted for in each sample is presented in Table 6.

- - - - - - - - - - - - - - -

Insert Table 6 here

- - - - - - - - - - - - - - -

The mean number of items, rounded to the nearest whole number, remaining after this final deletion was 36. The average number of components retained was 5, accounting for a mean of 48.2% of the variance in the remaining datasets. This is somewhat higher than the results reported in previous studies (Mitchell & Klimoski, 1982).

The mean interfactor correlations for the twelve samples are presented in Table 6. Due to the fact that the correlations were low; the highest correlation was .24, a varimax (orthogonal) rotation was used to obtain the final factor pattern matrices. Variation existed, across samples, in the particular components that were retained. Table 7 displays the labels that were given to each retained component and a brief description of them.

- - - - - - - - - - - - - - -

Insert Table 7 here

- - - - - - - - - - - - - - -

The components as retained for each dataset are presented in Table 8. The number of items loading on each factor and the percentage of variance in each dataset accounted for by these components is also presented.

- - - - - - - - - - - - - - -

Insert Table 8 here

- - - - - - - - - - - - - - -

A component scoring matrix was computed using the regression method as provided in *SPSS 7.0* (Norusis, 1994). This matrix was used to compute component scores for each individual for the retained components. Responses to the cross-validation samples were standardized using the means and standard deviations from the developmental samples. Missing data were replaced with item means.

Multidimensional Item Response Theory (MIRT) key

Multidimensional IRT combines features from unidimensional IRT and factor analysis to model responses to items as a function of a number of dimensional constructs

(θ's) as opposed to one underlying latent trait (θ). This was accomplished by utilizing the *Mplus 2* (Muthen & Muthen, 2001). To date, this is one of the few commercially available programs that model the two characteristics of current biodata measurement standards. The multidimensional aspect of many biodata items could be modeled with other programs (i.e., *Testfact 2* – Wilson, Wood, & Gibbons, 1991) that are based on multidimensional conceptions of the common binary response model; however, since ideal biodata items (Mumford & Owens, 1987; Owens, 1976) offer graded response continua, these programs are less than ideal. The *Parscale 3* (Muraki & Bock, 1997) will deal effectively with graded response models (Samejima, 1969, 1972); however, requires, the unidimensionality assumption. *Polyfact 1* (Muraki, 1998) handles both of these situations; however, the program in its current iteration does not provide for the scoring of individuals. *Mplus 2* (Muthen & Muthen, 2001) provides for measurement and structural modeling (i.e., structural equation modeling), while allowing items to be explicitly binary or ordinal categorical (a characteristic of IRT modeling). The program also computes individual latent trait scores based on the dimensions retained in the model.

As the requirements for the dataset in MIRT modeling are very similar to those of the standard factor analytic techniques, the datasets were modified identically to those in the factor analytic keying. As before, Table 4 presents a summary of the number of items remaining for analysis after deletion due to missing data and response invariance.

As with the factor analytic procedures, the next step in the analysis involved creation of a bivariate correlation matrix of the remaining variables. Due to the fact that the modeling procedures in MIRT operate under the assumption that responses to items

are categorical in nature, and specifically ordered categorical for this data, bivariate polychoric correlations were computed. As before, variables that produced no correlations higher than .30 with any other variable were eliminated from further analysis. Table 9 presents the number of items remaining in each dataset this step. Across all samples the mean (rounded to the nearest whole number) number of items across all datasets was 117, which represented 39.5% of the total itemset remaining for further analysis.

- - - - - - - - - - - - - - -

Insert Table 9 here

- - - - - - - - - - - - - - -

Utilizing *Mplus* 2 (Muthen & Muthen, 2001), the matrices of remaining variables for each developmental sample was subjected to a preliminary factor analysis using a weighted least squares (WLS) estimation procedure. Unlike principal components, matrix inversion is performed, requiring positive definiteness from the data matrix. Non-positive definiteness renders a matrix unfactorable, leading the researcher to reexamine the matrix and its contents. As none of the datasets at this point proved to be positive definite, more stringent criteria on item retention were applied. After matrix reexamination, variables demonstrating no correlations of at least .40 with any other variable were deleted. Following this deletion, further factoring revealed that all but four of the datasets produced positive definite polychoric correlation matrices. For two of these (where the original samples sizes were n = 500), deleting variables that did not exhibit any correlations of at least .45 rendered positive definiteness. For the remaining two (initial n = 100) deletion of variables was performed incrementally until only

variables that demonstrated correlations of at least .60 remained. Table 10 presents the number of items remaining in each dataset following this phase.

- - - - - - - - - - - - - - -

Insert Table 10 here

- - - - - - - - - - - - - - -

Across all samples, the mean number of items (rounded to the nearest whole number) was 79, which represented 26.7% of the original itemsets remaining for further analysis. Following this step, missing values were replaced with item medians (rounded to the whole number).

After ensuring matrix positive definiteness in all datasets, the remaining items were subjected WLS factor analysis. Eigenvalue plots were used initially to estimate the number of latent dimensions underlying each sample set. Next a model testing framework was established by specifying factor solutions with the number of underlying factors ranging from –3 to +3 of the number of dimensions arrived at via scree analyses. Since a primary focus during this phase was to reduce the itemsets to a manageable level, the models were evaluated using a combination of subjective interpretation of the various model solutions, and visual inspection of the factor loadings. Items that did not load at least .55 (Tabachnik and Fidell, 1996, refer to this magnitude as a "good" loading) with any interpretable factor were deleted. The remaining datasets were factored successively in this manner, until the point was reached where the itemsets could be reduced no further. Final dataset dimension numbers were determined statistically using a combination of the chi-square difference test, root mean square error of approximation (RMSEA), and the root mean squared residuals (RMSR). Six of the datasets were

unfactorable due to failure to reach convergence; even with liberal convergence criteria. The resulting datasets, including final item number and number of dimensions per sample are presented in Table 11. Also included is the percent variance accounted for by the retained dimensions and the mean interdimension correlations.

- - - - - - - - - - - - - - -

Insert Table 11 here

- - - - - - - - - - - - - - -

The mean number of items, rounded to the nearest whole number, after this phase was 28, representing 9.5 % of the original itemsets. In all datasets, five dimensions were retained. The mean percent of variance in the remaining datasets explained by the retained number of dimensions was 64.8%. Though the mean interdimension correlation was low (*.16*), a great deal of variation existed in the matrices (with all resulting matrices having correlations that ranged from near zero to .4 or .45, a promax (oblique) rotation was utilized to arrive at the final dimension (factor) pattern matrices. As the resulting dimensions were qualitatively similar to those obtained in the scales developed via principal components, the labels displayed in Table 7 were used. The latent dimensions retained for each dataset are presented in Table 12. The number of items comprising each dimension and the percentage of variance accounted for by each is also presented.

- - - - - - - - - - - - - - -

Insert Table 12 here

- - - - - - - - - - - - - - -

Dimension scores were computed in the developmental samples using the score estimation procedure in *Mplus* 2 (Muthen & Muthen, 2001) for ordered categorical

measured variables. Model parameter estimates; including item thresholds, dimension loadings/item discrimination indices, interdimension correlations, and dimension variances, from the developmental samples were used to constrain the scoring routines in the cross-validation samples (B.O. Muthen, personal communication, June 16, 2001) to estimate dimension scores.

Validation and Cross-Validation

For the groups where FAA Academy performance was being considered, validation and cross-validation was performed using linear and logistic regression analyses. With the former, in the developmental sample, the FAA Academy final score was regressed on the transmuted composite score from the ATCS entrance exam. Next, the empirically derived key score, the factor analytic scores, and the MIRT scores were added successfully, in a step-up hierarchical fashion (Lautenschlager & Mendoza, 1987), to the initial model. Predicted FAA Academy scores were computed using the empirically derived regression weights from the four models. Finally, the predicted scores were correlated with the actual scores. Similarly, using the regression weights derived in the developmental samples, predicted FAA Academy scores were calculated in the cross-validation samples. As with the developmental samples, these predicted scores were correlated with the actual scores.

The amount of shrinkage for the validity coefficients calculated on each sample was estimated by comparing the developmental and cross-validation coefficients using a test of the differences between correlation coefficients obtained on separate samples (Cohen & Cohen, 1983?). Validity coefficients estimated across methods of estimating regression-based predicted FAA Academy scores (i.e., different biodata methods) were

compared with each other using a test for dependent correlation coefficients (i.e., those calculated on the same datasets) (Cohen & Cohen, 1983?).

An alternative series of analyses were performed on the same samples, with the actual FAA Academy pass/fail status defined as the dependent variables, rather than the score. Using logistic regression analyses (Hosmer & Lemeshow, 1989), pass/fail status was regressed on the transmuted composite ATCS entrance exam score, and the three biodata keys in the same manner as the previous analyses. As in the previous analyses, differences between developmental and cross-validation samples were evaluated using a test for independent samples (Cohen & Cohen, 1983?). Dependent samples analyses (Cohen & Cohen, 1983?) were performed to compare the validity coefficients between different methods of obtaining predicted probability of success. In order to evaluate the classification accuracy of the predicted probabilities (and pass/fail groups) following the logistic regression analyses, classification tables comparing predicted group membership to actual group membership were created using the /CLASSPLOT subcommand (of the LOGISTIC REGRESSION command) in *SPSS 7* (Norusis, 1994).

In the samples that incorporated FAA air traffic facility OJT training performance as a criterion, logistic regression analyses were conducted in the same manner as that from the previous analyses. Field performance status (pass/fail) was regressed successively on a) the transmuted composite ATCS entrance exam score, b) the FAA Academy final score, and c) the three separately derived biodata keys. Statistical tests were performed in accord with that previously mentioned to assess the amount of shrinkage between developmental and cross-validation samples, and the actual difference between validity coefficients calculated using different models for predicting pass/fail.

Classification tables were used to evaluate each model's ability to accurately assign individuals to pass or fail groups.

CHAPTER 4 - RESULTS

FAA Academy

Descriptive Information

Descriptive characteristics for each sample are presented in Tables 13 through 19. Using a 2:1 split, subjects were randomly assigned to developmental or cross-validation samples, respectively. For all developmental samples (n=7), the average proportion of individuals passing to those failing the FAA Academy program was 61.1/38.9, roughly 3/2. For cross-validation samples this proportion was 59.2/40.8. For final Academy scores the average score mean and standard deviation, respectively, were 71.2 and 11.5. For the cross-validation samples these averages were 71.5 and 11.4. For ATCS entrance exam scores, the developmental sample average score mean and standard deviation were 90.7 and 4.7; while those for the cross-validation samples were 90.6 and 4.7. Similarly, the biodata score developmental sample to cross-validation sample comparisons yielded a high degree of homogeneity between samples.

- - - - - - - - - - - - - - -

Insert Table 13 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 14 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 15 here

- - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 16 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 17 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 18 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 19 here

- - - - - - - - - - - - - - -

Intercorrelations between the variables used in the analysis for the developmental samples are presented in Tables 20 through 26. Numbers in the lower diagonal represent those that were not corrected for range restriction, while those in the upper diagonal have been corrected for direct and indirect range restriction (Ghiselli, Campbell, & Zedeck, 1981). Of note across all samples is that there is a consistently significant relationship (uncorrected mean r = .25, corrected mean r =.57) between FAA Academy performance (as represented by final score) and the ATCS entrance exam score. Additionally a consistently significant relationship (uncorrected mean r = .31, corrected mean r =.44) exists between Academy performance and the empirically keyed biodata score.

- - - - - - - - - - - - - - -

Insert Table 20 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 21 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 22 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 23 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 24 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 25 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 26 here

- - - - - - - - - - - - - - -

Regarding the scale scores estimated via latent trait methods (factor analysis and multidimensional item response theory), and consistently significant relationship existed between that factor/dimension labeled "General Expectations" and Academy performance (FA: uncorrected mean r = -.18, corrected mean r =-.25; MIRT: uncorrected mean r = -.15, corrected mean r =-.25). A similar relationship exists between the factor/dimension labeled "High School Academics" and Academy performance (FA: uncorrected mean r = .14, corrected mean r =.26; MIRT: uncorrected mean r = .15, corrected mean r =.32). "Previous Air Traffic/Military Experience" displayed results that tended toward nonsignificance in uncorrected coefficients (FA: r = .07; MIRT: r = -.01), but upon correction this pattern changed somewhat (FA: r = .16; MIRT: r =- .10). Correlations between "Work Ethic/Personal Characteristics" and Academy performance remained low (FA: uncorrected mean r = .06, corrected mean r =.08; MIRT: uncorrected mean r = .07, corrected mean r =.07), as did those of Academy performance with "High School Athletics" (FA only) (uncorrected mean r = -.08, corrected mean r =-.07) and "Job Security as motivator of job choice" (MIRT only) (uncorrected mean r = -.10, corrected mean r =-.14).

Some other relationships of note are the generally significant coefficients between the ATCS entrance exam and empirical biodata key.(mean r =.16). Two of the latent dimensions that exhibited notable relationships with the exam score were "High School Academics" (FA: mean r =.11; MIRT: mean r =.15) and "Previous Air Traffic/Military Experience" (FA: mean r =-.19; MIRT: mean r =-.19). Additionally, there was a demonstrably significant relationship between the empirically derived key and the "High School Academics" scores (FA: mean r =.50 ; MIRT: mean r =.54) and the MIRT derived

"Previous Air Traffic/Military Experience" (mean r =-.20). Finally, it is noteworthy, and encouraging, that on average, the correlation between analogous latent trait scales developed via the two different methods (FA and MIRT) was r =.88.

Regression Analyses

In order to establish an empirical link between predictor (ATCS exam and biodata scores) and criterion (FAA Academy performance); and an optimal set of scoring weights, a series of regression analyses were performed. Hierarchical linear regression was used for predicting Academy final scores from the combination of ATCS exam scores and biodata. Parallel analyses were done utilizing correlation matrices uncorrected and corrected for range restriction. Redefining Academy performance as pass/fail status, hierarchical logistic regression was used to develop the weights for optimally predicting individual probability of passing.

Predicting Academy Score

Results from regression analyses, using uncorrected correlations, for each developmental sample are presented in Tables 27 through 33. Across all samples the mean R^2 value was .07 when the only predictor was ATCS exam score. Minus the most extreme value (that for the smallest sample size), this value was .05. With the addition of the empirically derived biodata key, the mean R^2 value was .16 (this fell to .10 when the value from the smallest dataset was excluded). When the factorially derived biodata scores were added to the ATCS exam score, the mean R^2 value was .13 (.11 minus the smallest dataset). Finally, the addition of the MIRT-derived scores produced a mean R^2 value of .10.

- - - - - - - - - - - - - - -

Insert Table 27 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 28 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 29 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 30 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 31 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 32 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 33 here

- - - - - - - - - - - - - - -

Considering the contribution of individual variables, the average β for ATCS exam score when considered alone was .25. When taken in concert with the empirically

derived biodata score, this number drops to .19, while the average β for the biographical key was .28. For the analyses using the factorially created key, the average β for ATCS exam score was .24. The average β for the individual factor scores were: a) "General Expectations", -.17; b) "High School Academics", .12; c) "Previous Air Traffic/Military Experience", .09; d) "Work Ethic/Personal Characteristics", .06; e) "High School Athletics", -.07; and "Extracurricular Activity Involvement" (present in only one analysis), .06. Similarly, the average β for the exam score, in conjunction with the MIRT key, was .22. Individual dimension score average β's were: a) "General Expectations", -.13; b) "High School Academics", .14; c) "Previous Air Traffic/Military Experience", .11; d) "Work Ethic/Personal Characteristics", .06; and e) "Job Security Motivation", -.04.

Analogous results from the regression analyses that were performed using correlations that had been corrected for direct range restriction on the ATCS exam score and indirect range restriction on the biodata scale scores are presented in Tables 34 through 40. Across all samples the mean R^2 value was .34 when the only predictor was ATCS exam score. Leaving out the most extreme value (R^2 = .66 for the smallest sample), the mean value was .29. With the addition of the empirically derived biodata key, the average R^2 was .47 (mean R^2 = .38 when most extreme value, R^2 = 1.00, is removed). Adding the factor analytic scale to the ATCS exam score, yields an average R^2 of .46 (mean R^2 = .39 when most extreme value, R^2 = .86, is removed). Finally, the addition of the MIRT-derived scores produced a mean R^2 value of .42.

- - - - - - - - - - - - - - -

Insert Table 34 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 35 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 36 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 37 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 38 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 39 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 40 here

- - - - - - - - - - - - - - -

Considering the contribution of individual variables, the average β for ATCS exam score when considered alone was .57. When taken in concert with the empirically

derived biodata score, this number drops to .51, while the average β for the biographical key was .35. For the analyses using the factorially created key, the average β for ATCS exam score was .50. The average β for the individual factor scores were: a) "General Expectations", -.21; b) "High School Academics", .20; c) "Previous Air Traffic/Military Experience", -.02 (-.08 when most extreme removed); d) "Work Ethic/Personal Characteristics", .07; e) "High School Athletics", -.06; and "Extracurricular Activity Involvement" (present in only one analysis), .10. Similarly, the average β for the exam score, in conjunction with the MIRT key, was .50. Individual dimension score average β's were: a) "General Expectations", -.19; b) "High School Academics", .20; c) "Previous Air Traffic/Military Experience", -.04; d) "Work Ethic/Personal Characteristics", .06; and e) "Job Security Motivation", -.02.

Predicting Academy Pass/Fail Status

Results from logistic regression analyses for each developmental sample are presented in Tables 41 through 47. Across all samples the mean Nagelkerke (1991) R^2 value was .06 when the only predictor was ATCS exam score. Minus the most extreme value (that for the smallest sample size), this value was .05. With the addition of the empirically derived biodata key, the mean Nagelkerke (1991) R^2 value was .23 (this fell to .11 when the value from the smallest dataset was excluded). When the factorially derived biodata scores were added to the ATCS exam score, the mean Nagelkerke (1991) R^2 value was .12 (.10 minus the smallest dataset). Finally, the addition of the MIRT-derived scores produced a mean Nagelkerke (1991) R^2 value of .08.

- - - - - - - - - - - - - - -

Insert Table 41 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 42 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 43 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 44 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 45 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 46 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 47 here

- - - - - - - - - - - - - - -

Considering the contribution of individual variables, the average e^{β} for ATCS exam score when considered alone was 1.12. When taken in concert with the empirically

derived biodata score, this number was 1.31 (1.08 minus the extreme value) , while the average e^{β} for the biographical key was 2.04 (1.99 minus the extreme value). For the analyses using the factorially created key, the average e^{β} for ATCS exam score was 1.10. The average e^{β} for the individual factor scores were: a) "General Expectations", .72; b) "High School Academics", 1.21; c) "Previous Air Traffic/Military Experience", 1.15; d) "Work Ethic/Personal Characteristics", 1.17; e) "High School Athletics", .89; and "Extracurricular Activity Involvement" (present in only one analysis), 1.16. Similarly, the average e^{β} for the exam score, in conjunction with the MIRT key, was 1.09. Individual dimension score average e^{β}'s were: a) "General Expectations", .74; b) "High School Academics", 1.34; c) "Previous Air Traffic/Military Experience", 1.39; d) "Work Ethic/Personal Characteristics", 1.23; and e) "Job Security Motivation", .89.

Validation/Cross-Validation

Prediction of Academy Score

Predicted criterion performance was correlated with observed performance in both the developmental and cross-validation samples (weights derived in the developmental samples were applied to the cross-validation samples). The predictive models were compared with regard to a) validity coefficient shrinkage, b) cross-validity coefficient magnitude and significance, and c) statistically significant differences between keying procedures. Table 48 presents validities and cross-validities as a function of keying method and sample size (based on uncorrected information) for the seven samples.

- - - - - - - - - - - - - - -

Insert Table 48 here

- - - - - - - - - - - - - - -

Developmental and cross-validation sample correlation coefficients were compared using a test for the difference between independent sample statistics (Cohen & Cohen, 1983, p. 53). At the largest sample sizes (dev. n = 3,114, c.v. n = 1,454) no shrinkage in validity coefficients was observed. In fact, the coefficients for two of the keyed samples (no biodata and the empirical key) actually increased (from .20 to .26, and .29 to .31, respectively). Those for the other two (factor analytic, .32; and MIRT, .30) remained the same from developmental to cross-validation samples. For the largest sample, none of the differences that did exist was statistically significant. Though some shrinkage occurred for most keying methods for the next three (original N = 3,000; 2,000; or 1,500) samples, none of these differences was statistically significant.

At an original sample size of 1,000, no MIRT key was available, so comparisons at this point did not include this. For all samples at this sample size, differences that did exist were non-significant. With an original developmental sample of n = 500, the differences between developmental and cross-validation samples were not significant for the "No Biodata" procedure, but were for the empirical (z=3.50, $p \leq .01$) and factor analytic (z=2.70, $p \leq .01$) key enhanced predictive models. Similarly, though no significant differences were found for the "No Biodata" model at the smallest (n=100) sample, the differences for the empirical (z=3.03, $p \leq .01$) and factor analytic (z=2.15, $p \leq .05$) models were significant.

From the perspective of observing a noticeable trend, there were no developmental to cross-validational coefficient differences that appeared to be present. When the total samples dipped to N=500, these differences, whether significant or not, were magnified considerably. From the perspective of the magnitude of the obtained

z-value (for those differences that were significant), those obtained for the model with empirical key (z=3.50 and 3.03) were greater than those obtained for the factor analytic key enhanced model (z=2.70 and 2.15).

Table 49 shows the coefficients from the cross-validation samples only. Presenting these coefficients in this manner allows for a more readily available comparison based on keying procedure and sample size differences. Concerning differences in sample size, the addition of the factor analytic key consistently produced the highest coefficients in the four largest samples (where keys for all procedures existed), with a mean coefficient of $r = .33$. Following were the MIRT (mean $r = .33$), empirical (mean $r = .31$), and "No Biodata" (mean $r = .23$). This pattern was somewhat different at the smaller sample sizes; however, as the empirical key coefficients outperformed (mean $r = .20$) the factor analytic and "No Biodata" models (both with a mean $r = .15$). For the three procedures with complete data across all sample ranges, the empirical and factor analytic procedures yielded a mean $r = .25$, while that for the "No Biodata" procedure was $r = .19$. Generally, for all procedures the impact of sample size on validity coefficient is not of note until the total sample size falls below $N = 1{,}000$. However, all procedures produced cross-validities that were significant until the total sample size dipped to $N = 500$. Here and at $N = 100$ none of the cross-validities achieved significance.

- - - - - - - - - - - - - - -

Insert Table 49 here

- - - - - - - - - - - - - - -

To evaluate the differences between procedures, significance tests were conducted on the coefficients from the cross-validation samples. For each sample size, the coefficient from each procedure was compared to the others for a particular sample size (see Table 49). For each of the four largest samples, six pairs of validity coefficients were tested for differences in procedures, using a *t*-test for comparison of dependent samples (Cohen & Cohen, 1983, p. 56). Across the seven samples, no significant differences between methods were found for the cross-validities. Though not explicitly part of the analysis, the same test conducted on the developmental samples also yielded no significant differences between methods.

Table 50 presents validities and cross-validities as a function of keying method and sample size (based on corrected information) for the seven samples. At the largest sample sizes (dev. n = 3,114, c.v. n = 1,454) no shrinkage in validity coefficients was observed. In fact, the coefficients for three of the keyed samples (no biodata, empirical key, and factor analytic key) actually increased (from .20 to .26; .28 to .32; and .28 to .29 respectively). Those for the MIRT key, coefficients (r=.27) remained the same from developmental to cross-validation samples. For the largest sample, none of the differences that did exist was statistically significant. Though some shrinkage occurred for all keying methods for the next three (original N = 3,000; 2,000; or 1,500) samples, none of these differences was statistically significant.

- - - - - - - - - - - - - - -

Insert Table 50 here

- - - - - - - - - - - - - - -

For all available samples at N=1,000, differences that did exist were non-significant; however, rather than cross-validity shrinkage the pattern was for the coefficients to be inflated between the two samples. With an original developmental sample of n = 500, the differences between developmental and cross-validation samples were not significant for the "No Biodata" or "Factor Analytic" procedures, but was for the empirical (z=2.74, $p \leq .01$) key enhanced predictive model. This pattern held at the smallest sample (n=100) sample, as no statistically significant difference existed between developmental and cross-validation samples for the "No Biodata" and "Factor Analytic" models, but did for the empirical key (z=3.00, $p \leq .01$). As before, no noticeable trend in the differences between developmental to cross-validation coefficients appeared to be present. When the total samples dipped to N=500 and below, these differences, whether significant or not, were magnified considerably.

Table 51 shows the coefficients from the cross-validation samples for the analyses using corrected data. Concerning differences in sample size, the addition of the factor analytic key did not produce the highest coefficients as it had in the uncorrected data. Averaging across the four largest samples (where all methods were observed) the mean cross-validities were highest for the empirical key r = .29. Following closely were the factor analytic (mean r = .28), and MIRT (mean r =.27). The "No Biodata" model (mean r = .23) was, as before, the lowest. This pattern held up at the smaller sample sizes as the empirical key coefficients outperformed (mean r = .20) the factor analytic and "No Biodata" models (with mean r's of .17 and .15, respectively). For the three procedures with complete data across all sample ranges, the empirical and factor analytic procedures yielded mean r's of .25 and .23, respectively, while that for the "No Biodata" procedure

was $r = .19$. Generally, for all procedures the impact of sample size on validity coefficient is not of note until the total sample size falls below $N = 1{,}000$. However, all procedures produced cross-validities that were significant until the total sample size dipped to $N = 500$. Here and at $N = 100$ none of the cross-validities achieved significance.

- - - - - - - - - - - - - - -

Insert Table 51 here

- - - - - - - - - - - - - - -

Evaluating the differences between procedures via significance tests no significant differences between methods were found for the cross-validities, across the seven samples. As before (using uncorrected data) significance tests conducted on the developmental samples also yielded no significant differences between methods.

Prediction of Academy "Pass" or "Fail"

Table 52 presents comparative validity coefficients for developmental and cross-validation subsamples for all sample sizes, using Academy pass/fail as the criterion. At the largest sample size, one of the coefficients exhibited shrinkage (MIRT went from $r =$ .23 to .22), two showed no change at all (empirical and factor analytic: $r = .23$ and .24, respectively), and one ("No Biodata") demonstrated inflation from developmental ($r =$ .16) to cross-validation ($r = .20$) sample. None of the developmental to cross-validation differences was statistically significant. Though some shrinkage occurred for most keying methods for the next three samples (original $N = 3{,}000$; 2,000; or 1,500), these differences were not statistically significant.

- - - - - - - - - - - - - - -

Insert Table 52 here

- - - - - - - - - - - - - - -

At a sample size of N = 1,000, three different patterns emerged for the three available samples (all but the MIRT were present). The factor analytic scale exhibited shrinkage, the empirical showed no change, and the "No Biodata" model produced an inflated cross-validity. None of the differences that existed was statistically significant. At the two smallest sample sizes (N = 500 and 100) all of the cross-validities demonstrated shrinkage. Though this shrinkage proved to be non-significant for the "No Biodata" model for both samples, that for the empirical (N = 500, z = 4.33, $p \leq .01$; N = 100, z = 8.53, $p \leq .01$), and factor analytic (N = 500, z = 2.62, $p \leq .01$) models. The shrinkage exhibited at the N = 100 sample using s factor analytic model was not significant. As with the previous comparisons, utilizing the Academy score as a criterion, no recognizable pattern in cross-validity shrinkages occurred until the smaller sample sizes ($N \leq 500$). Shrinkage was greatest for the model including the empirical key.

Table 53 shows the coefficients from the cross-validation only. As before, the addition of the factor analytic key to the ATCS exam score consistently produced the highest cross-validities in the four largest samples (mean r = .26). Following were the MIRT (mean r = .24), empirical (mean r = .23), and the "No Biodata" (mean r = .18). At the three smallest samples, the mean r for the empirical method was .15, while the factor analytic and "No Biodata" both exhibited mean r's of .15. For the three procedures with complete data across all sample ranges, the factor analytic model (mean r = .21) outperformed the empirical (mean r = .19) and "No Biodata" (mean r = .16) models.

Generally, the impact of validation sample size is negligible across all samples until $N<$ 1,000, with the most drastic drops occurring at the N=500 sample. All cross-validities from the 4,568 to 1,000 samples displayed statistical significance; whereas, none at the two smallest samples achieved such.

- - - - - - - - - - - - - - -

Insert Table 53 here

- - - - - - - - - - - - - - -

No statistically significant differences between the cross-validities obtained via different methods were found. Of interest, however, is the fact that at the N = 100 sample, the validities for the developmental sample comparisons between the "No Biodata" and empirical key models ($t_{(60)} = -2.05$, $p \leq .05$) and the empirical and factor analytic models ($t_{(60)} = 2.04$, $p \leq .05$) were statistically significant.

Classification accuracy. The logistically derived predictive models for Academy pass or fail were evaluated using a contingency table based classification model. Percentage of correct classifications in the cross-validation samples (to pass or fail status) are calculated and used as the standard of comparison. These percentages for each sample size are presented in Table 54. For the comparison between methods when all models were present (N = 4,568 to 1,500) the models including the biodata keys outperformed the ATCS exam only model (mean % correct = 59.7). The factor analytic key model (mean % correct = 62.7) outperformed both the empirical (mean % correct = 62.5) and MIRT (mean % correct = 62.3) models. When considering comparisons across all seven samples, the factor analytic approach (mean % correct = 61.5) outperformed the empirical (mean % correct = 60.8) and "No Biodata" (mean % correct = 59.6) models.

- - - - - - - - - - - - - - -

Insert Table 54 here

- - - - - - - - - - - - - - -

FAA Field Facility Performance

Descriptive Information

Descriptive characteristics for each sample are presented in Tables 55 to 59. As before, a 2:1 (developmental/cross-validation) sample split was conducted. For all developmental samples, the average proportion of individuals passing (reaching Full Performance Level) to those failing (facility wash-out) was 90.4/9.6, roughly 9/1. For cross-validation samples, this proportion was 91.4/8.6. Regarding final FAA Academy performance, the average mean and standard deviation for scores (for the developmental sample) were 79.0 and 6.2, respectively. For the cross-validation samples, these were 78.8 and 5.8, respectively. For ATCS entrance exam scores, the developmental sample average score mean and standard deviation were 91.2 and 4.8; while those for the cross-validation samples were 90.9 and 4.9. Similarly, the biodata score developmental sample to cross-validation sample comparisons yielded a high degree of homogeneity between samples.

- - - - - - - - - - - - - - -

Insert Table 55 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 56 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 57 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 58 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 59 here

- - - - - - - - - - - - - - -

Intercorrelations between the variables used in the analysis for the developmental samples are presented in Tables 60 through 64. Numbers in the lower diagonal represent coefficients that have not been corrected for restriction in range. Of note, across all samples, is that there is a consistently low and non-significant relationship between On-the-job (OJT) FAA field facility performance (as represented by attainment of FPL certification or "wash out") and the FAA Academy score (mean absolute r = .06) or ATCS entrance exam score (mean r =-.09). However, of equal note is the consistently high and significant relationship between facility performance and the empirically keyed biodata score (mean r =.36, with coefficients ranging from .18 to .79).

- - - - - - - - - - - - - - -

Insert Table 60 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 61 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 62 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 63 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 64 here

- - - - - - - - - - - - - - -

Regarding the scale scores estimated via latent trait methods (factor analysis and multidimensional item response theory), a relationship existed between that factor/dimension labeled "General Expectations" and facility performance (FA: mean r = .12; MIRT: mean r = .12). The relationships between facility performance and other latent dimensions were: "High School Academics" (FA: mean r = -.05; MIRT: mean r = -.05); "Previous Air Traffic/Military Experience" (FA: mean r = .06; MIRT: mean r = .05); "Work Ethic/Personal Characteristics" (FA: mean r = .08; MIRT: mean r = .05); "High School Athletics" (FA only) (mean r = .10); "Job Security as motivator of job

choice" (mainly MIRT – showed up in one of the FA, r = -.02) (mean r = .08). Two other latent dimensions that appeared sporadically in the factor analytically developed keys were "Attendance" (mean r with facility performance is .06) and "Interpersonal Affiliation" (once, r = .26, in the N=100 sample).

Some other relationships of note are the generally significant coefficients between FAA Academy final score with the ATCS entrance exam (mean r =.14) and "High School Academics" (FA: mean r = .10; MIRT: mean r = .11). Two of the latent dimensions that exhibited notable relationships with the exam score were "High School Academics" (FA: mean r =.15; MIRT: mean r =.16) and "Previous Air Traffic/Military Experience" (FA: mean r =-.21; MIRT: mean r =-.21). Unlike the empirical biodata keys from the previous set of analyses, no consistent across-the-board relationships with the latent dimensions were found. There were, however, some compelling correlations with "General Expectations" (FA: mean r =.23 ; MIRT: mean r =.22); "Work Ethic" (FA: mean r =.20 ; MIRT: mean r =.10); "Previous Air Traffic/Military Experience" (FA: mean r =.28 ; MIRT: mean r =.25); "High School Academics" (FA: mean r =-.15; MIRT: mean r =-.14); and "High School Athletics" (FA only) (mean r =.11). Finally, as in the previous set of analyses, it is encouraging, that on average, the correlation between analogous latent trait scales developed via the two different methods (FA and MIRT) was r =.89.

Regression Analyses

In order to establish an empirical link between predictor (ATCS exam, FAA Academy performance, and biodata scores) and criterion (FAA field facility performance); and an optimal set of scoring weights, a series of regression analyses were

performed. Hierarchical logistic regression was used for predicting facility pass/fail status from the combination of ATCS exam and FAA Academy scores and biodata.

Predicting Field Facility "Pass/Fail" Status

Results from logistic regression analyses for each developmental sample are presented in Tables 65 through 69. Across all samples the mean Nagelkerke (1991) R^2 value was .04 when the only predictors were ATCS exam score and FAA Academy performance. Minus the most extreme value (that for the smallest sample size), this value was .02. With the addition of the empirically derived biodata key, the mean Nagelkerke (1991) R^2 value was .18 (this mean was representative of four samples, as nonconvergence of the logistic solution for the smallest sample prevented sample statistics). When the factorially derived biodata scores were added to the ATCS exam score, the mean Nagelkerke (1991) R^2 value was .15 (.09 minus the smallest dataset). Finally, the addition of the MIRT-derived scores produced a mean Nagelkerke (1991) R^2 value of .05.

- - - - - - - - - - - - - - -

Insert Table 65 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 66 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 67 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 68 here

- - - - - - - - - - - - - - -

- - - - - - - - - - - - - - -

Insert Table 69 here

- - - - - - - - - - - - - - -

Considering the contribution of individual variables, the average e^{β}'s for ATCS exam score and Academy performance were considered alone were .94 and 1.01, respectively. When taken in concert with the empirically derived biodata score, this numbers were .98 and 1.03, while the average e^{β} for the biographical key was 1.32. For the analyses using the factorially created key, the average e^{β}'s for ATCS exam score and Academy performance were .97 and 1.00. The average e^{β}'s for the individual factor scores were: a) "General Expectations", 1.65; b) "High School Academics", .75; c) "Previous Air Traffic/Military Experience", 1.26; d) "Work Ethic/Personal Characteristics", 1.42; e) "High School Athletics", 1.43; f) "Attendance" (present in only two analyses), 1.22; g) "Job Security Motivation" (only one analysis), .92; and h) "Interpersonal Affiliation" (present only in analysis of $N = 100$), 3.28. Similarly, the average e^{β} for the exam and Academy scores, in conjunction with the MIRT key, were .97 and 1.04. Individual dimension score average e^{β}'s were: a) "General Expectations", 1.64; b) "High School Academics", .83; c) "Previous Air Traffic/Military Experience", 1.11; d) "Work Ethic/Personal Characteristics", 1.19; and e) "Job Security Motivation", 1.20.

Validation/Cross-Validation

Prediction of Field Facility "Pass" or "Fail"

Table 70 presents comparative validity coefficients for developmental and cross-validation subsamples for all sample sizes, using field facility pass/fail as the criterion. At the largest sample size (N = 2,000), one of the coefficients exhibited shrinkage (the empirical key enhance sample went from r = .17 to .06), one showed no change at all (factor analytic: r = .17), and two (MIRT and "No Biodata") demonstrated inflation from developmental (r = .15 and .06, respectively) to cross-validation (r = .17 and .13, respectively) sample. Of those differences that did exist, only the one exhibiting shrinkage was statistically significant (z=2.48, $p \leq .05$). At the N = 1,500 sample, all of the cross-validation subsamples exhibited shrinkage. Of these, the empirical (z=4.12, $p \leq .01$) and factor analytic (z=2.18, $p \leq .05$) key models displayed statistical significance, while the others did not.

- - - - - - - - - - - - - - -

Insert Table 70 here

- - - - - - - - - - - - - - -

At a sample size of N = 1,000, two different patterns emerged for the three available samples (all but the MIRT were present). The "No Biodata" model exhibited inflation, and the empirical and factor analytic models displayed shrunken cross-validities. Of the two latter, only the empirical model produced statistically significant shrinkage (z=4.51, $p \leq .01$). At the two smallest sample sizes (N = 500 and 100) all of the cross-validities demonstrated shrinkage. Though this shrinkage proved to be non-significant for the "No Biodata" model for the N = 500 sample, those for the empirical (z

= 6.68, $p \leq .01$), and factor analytic ($z = 4.28$, $p \leq .01$) models were. The shrinkage exhibited at the $N = 100$ sample using both a "No Biodata" ($z = 2.67$, $p \leq .01$) and factor analytic ($z = 3.65$, $p \leq .01$) model was significant.

Unlike previous comparisons, utilizing the Academy score as a criterion, a clear pattern in cross-validity shrinkages occurred when using facility pass/fail as a criterion. Generally speaking, the magnitude of the differences between validation and cross-validation samples tended to increase as sample size decreased. All of these comparisons were statistically significant for the model incorporating the empirical key, and those for the empirical key appeared to be the largest.

Table 71 shows the coefficients from the cross-validation only. The addition of the factor analytic key to the ATCS exam and FAA Academy score consistently produced the highest cross-validities in the three largest samples (mean $r = .15$). Following were the MIRT (mean $r = .13$, based on two samples), "No Biodata" (mean $r = .10$), and the empirical (mean $r = .04$). At the two smallest samples, the mean r for the factor analytic method was r = -.18, while the mean for the "No Biodata" model was r = -.15. Based on the one available coefficient (for the $N = 500$) for the empirical model, a cross-validity of $r = -.08$ was produced. In Table 4c.04 it is clearly illustrated that cross-validity magnitudes decrease as sample size decreases. All cross-validities, but those for the empirical key, from the 2,000 to 1,000 samples displayed statistical significance; whereas, none at the two smallest samples achieved such. No statistically significant differences between the cross-validities obtained via different methods were found.

- - - - - - - - - - - - - - -

Insert Table 71 here

- - - - - - - - - - - - - - -

Classification accuracy. The logistically derived predictive models for field facility pass or fail were evaluated using a contingency table based classification model. Percentage of correct classifications in the cross-validation samples (to pass or fail status) are calculated and used as the standard of comparison. These percentages for each sample size are presented in Table 72. For the comparison between methods when all models were present (N = 2,000 to 1,500) all models performed equally well at correctly classifying individuals (mean % correct = 90.5). At the N = 1,000 and 500 samples (no MIRT key was available), the "No Biodata" and factor analytic methods on average correctly classified 92.8% of all individuals, while the empirical method trailed at 91.6%. At the N = 100, since no empirical keyed model was present the comparison between the remaining two yielded a correct classification level of 90.0% for the "No Biodata" and 87.9% for the factor analytic model.

- - - - - - - - - - - - - - -

Insert Table 72 here

- - - - - - - - - - - - - - -

CHAPTER 5 - DISCUSSION

What is the impact of item scaling method on incremental validities and cross-validities of a biographical inventory for predicting performance in school or on the job? What effect do the size of the developmental and cross-validation samples have on the stability of the different key types? In this paper, these questions were addressed via empirical comparisons of criterion prediction models developed on samples of air traffic control students that attended the FAA Academy between 1988 and 1990. Biographical information was scaled using three different procedures and the resulting information was combined with an ability-based selection battery to form scoring keys that were compared to each other for shrinkage and overall validity, when predicting Academy or subsequent OJT field performance.

In addition to these questions, the other important inquiry that was addressed pertained to one of the biodata scaling methods. In particular, multidimensional item response theory (MIRT) was applied to biodata items in an attempt to approach biographical information via modern measurement theory. As Ackerman (1996) has pointed out, MIRT is still a fairly recent addition in the measurement professional's toolkit, and applications of such are still rare in the literature. Comparison of scoring models utilizing MIRT with those developed using a classical "dustbowl empiricist" focus or standard linear factor analytic methods, helped to put this approach in perspective and pave the way for its further application.

Briefly, the methodology that was employed in this study consisted of : a) generating a number of different sample sizes from air traffic control students that attended the FAA Academy between 1988 and 1990; b) dividing each sample into

developmental and cross-validation components, using a 2:1 ratio; c) creating three different biodata keys on each developmental subsample, using biographical data collected with multiple-choice format questionnaires; d) applying the developmental keys to the cross-validation sub-samples; e) combining the biodata keys with air traffic control entrance exam scores (or exam scores and Academy performance scores); f) regressing Academy (or field) performance on the aforementioned predictors; g) applying the empirical regression weights to form predictor composites in the developmental and cross-validation samples; h) correlating each composite with actual Academy (or field) performance; in) comparing developmental and cross-validity coefficients for evidence of shrinkage and overall predictive power; and (for those composites developed using logistic regression) j) conducting classification analyses to investigate the ability to correctly assign individuals to "pass" or "fail" conditions.

In terms of the hypotheses of interest, the results of this study tended to lend modest support. The first hypothesis dealt directly with the magnitude between the developmental validities of the empirical biodata key and the factorially derived scales. It was expected that those of the empirical key would be higher than the factorial keys. At whole sample sizes of

N = 1,000 or less (with developmental sizes ranging from n = 663 to 63), this was case for the Academy only dataset when Academy score was used as the criterion (using both uncorrected and corrected data). When the criterion of interest became Academy pass/fail status, the hypothesized effects did not occur until the N = 500 or less comparisons. For the Academy-Field Facility comparisons, the differences between empirical and factor analytic initial validities was apparent at the N = 1,500 analysis. It was interesting to note

that at the larger sample sizes, differences between methods for initial validities were minimal, if they existed at all. Though the initial validities in this study did not approach the magnitude of those reported in Devlin, et. al. (1992), or Mitchell and Klimoski (1982), the findings in this study are consistent with theirs.

Whereas the first research hypothesis dealt with differences between methodologies that existed in the developmental samples, the second pertained to the cross-validation samples. Specifically, it was hypothesized that the differences between empirical and factor analytic keys would be minimal. For the most part this was the finding. For the prediction of Academy final score (using uncorrected or corrected correlations) average cross-validities across all samples for the two methods were quite comparable. This trend was duplicated in the analyses where the Academy performance criterion was "pass" or "fail" status. This was not the same pattern that was observed in the Academy-Field Facility data; however. Here, the addition of the factor analytic key to the selection test outperformed the empirical key consistently.

Finally, it was hypothesized that the magnitude of the differences between developmental and cross-validation coefficients would be greater at smaller sample sizes. This in fact was supported. Though the amount of shrinkage at the larger sample sizes (N>1,000) was for the most part non-existent for the Academy only samples, this changed dramatically at the smaller ones. For the Academy-Field Facility samples; however, the trend became apparent for almost all samples.

Though not explicitly included as a research group per se, the ATCS exam only comparisons mirrored those in the "exam + biodata key" groups; however, estimates tended to be more conservative. At larger samples, for the Academy only groups,

differences between validation and cross-validation samples were non-existent, though they did increase as sample sizes decreased. It was encouraging from this researcher's view point that coefficients were higher for those samples including biodata keys.

As mentioned earlier, there are few if any studies comparing validity and cross-validity evidence for multidimensional item response theoretic models. Since convergence in biodata key parameters at sample sizes below $N = 1{,}500$ was not reached, it was not possible to evaluate the stability of MIRT biodata keys in this study. However, for those samples where MIRT keys were available, it should be noted that estimates were very close to those obtained via the factor analytic methods. This is not surprising given that both methods produced multiple scale scores per key and were both roughly based around the same latent variable conceptualization. Though the results are by no means definitive, the results from this study indicated that the MIRT key; though not performing as well as the factor analytic analyses, outperformed the empirical keys.

Though the results indicate that each of the biodata keys added incremental predictive validity to the ATCS exam scores in predictive Academy performance, classification accuracy appeared to be not be greatly affected by the addition of the biodata keys. For the most port, each of the keys produced about a 1% bump up in classification effectiveness. This result was somewhat discouraging from a research perspective. From a practical significance stand point; however, the 1% bump could have equated to about 45 more people being classified accurately had the biodata screen been implemented. For the Academy-Field Facility samples, classification accuracy was not improved when the biodata keys were added to the ATCS exam and Academy scores as a predictor of facility certification or "was out." This apparent ineffectiveness of the added

biodata keys at the facility level may be a function of the fact that nearly 90+% of those in the samples had reached FPL status.

Outside of the research questions explicitly addressed in this study, some of the findings that came out the research pertained to the biodata keys and how they interacted with the exam score in the prediction of Academy performance. Generally, the empirical key contributed the most incrementally as a single element when combined with the ATCS exam scores in the developmental samples. This is entirely consistent with past research and did not provide any earthshaking enlightenment to the biodata scaling literature. Also, it is important to note that in the samples where the empirical key consisted of only a single item ($4{,}568 \geq N \geq 1{,}500$), this item was self reported high school math grades; which again is consistent with the literature (Collins, Nye, & Manning, 1990). This item also turned up as a component of the keys derived on smaller samples, as did that pertaining to self-expected relative performance. Collins, Nye, and Manning (1990) reported this latter finding as well.

For the latent dimensions, the contribution of the analogous domains estimated via principal components or MIRT were essentially the same. High school academic performance and previous air traffic/military experience contributed the most to the prediction of Academy performance. This is consistent with expectations as these are the two domains that would appear to follow the "signs" and "samples" guidelines established by Wernimont and Campbell (1968). General occupational expectations, which included items dealing with supposed job autonomy, working with others, and others pertaining to the job and work environment, demonstrated the strongest negative relationship with Academy performance. This was somewhat confusing from the

perspective that the manner in which the Academy screen program was run at the time, would have led this researcher to believe a positive relationship. This is so because it would seem that the Academy environment would be somewhat similar to the work setting and it would be expected that personal expectations would be a more direct predictor of eventual performance.

For the prediction of actual field performance, the empirical key again was the greatest single predictor of eventual certification. Here; however, no single item keys were produced. Items that showed up consistently at all sample sizes included those pertaining to self-reported work ethic and general social interactions. This was mirrored in the latent scales as the dimensions that contributed the most to prediction over and above entrance exam and Academy performance were those that pertained to: a) general work expectations, b) work ethic/personal characteristics, and c) high school athletic participation. All of these domains and the single items in the empirical key all represent areas that would be expected to contribute to the ability to work with others and at the same time be able to perform well in a job requiring fairly high level functioning. It is interesting to note that whereas high school academic (and in particular math grades) performance was a fairly decent predictor of Academy performance, it had an inverse effect on the prediction of field performance.

There are four major implications or contributions that this research has for the field at large. First it makes a needed contribution to the comparative biodata keying literature in the same vein as Devlin, et al. (1992) or Mitchell and Klimoski (1982). It has provided guidelines for appropriate sample sizes in validation research and has supported earlier findings (Mitchell & Klimoski, 1982) comparing empirical with factorial keying

methods. Second, it has provided an introduction of a new methodological technique to the applied literature. Utilization of IRT has been relatively underrepresented in measurement arenas outside of education (particularly I/O psychology) and ongoing published applications are necessary. Extending this contribution, Ackerman (1996) has pointed out that applications of MIRT are almost nonexistent, even in education. Hence, this study provides positive input to the literature. One last contribution pertains to the application of modern measurement theoretic modeling to the scaling and understanding of biographical data.

One limitation of this study is that though a number of different sample sizes were used, a greater number of samples and finer distinctions between these samples would have provided a better estimate of the validity "breaking point" for each method. Developmental to cross-validation comparisons demonstrated essentially no shrinkage until around a starting (prior to splitting into validational groups) sample size of 1,000. Of interest would be the exact point at which demonstrable shrinkage takes place.

Of particular interest to this researcher would have been a more detailed study involving a more diverse array of criterion measures. Though eventual Academy score or pass/fail, or field-level certification or certainly acceptable criterion, other criteria of interest would include: a) Academy over-the-shoulder rankings and/or laboratory exercise performance, b) field-level performance checks (including comparisons to other controllers), c) numbers of operational errors, d) time to reach full field-level performance checks (including comparisons to other controllers), c) numbers of operational errors, d) time to reach full performance level, and e) salary progression.

From a methodological standpoint, one weakness pertained to the use of the empirical key that was created using a "pass/fail" criterion, in a validation equation where Academy "score" was being predicted. Though in this study it appeared to have no noticeable effect on the results, the reduction in variance that the "pass/fail" keying strategy would tend to have could definitely have provided misleading results. Perhaps a more drastic weakness is that relating to the choice of the certification/no certification (characterized as pass/fail) at the facility level. With an overall pass rate of over 90%, the likelihood of failure can be considered a rare event. King and Zeng (2001) explain in a fair amount of detail how rare events can be quite problematic for standard regression procedures. In fact, as noted in this study, classification accuracy remained unchanged regardless of the predictor used to estimate the outcome.

Future research efforts should focus on directly addressing the aforementioned weaknesses. Though some (Mumford, 1996) doubt the efficacy of continuing research on comparing scaling biodata scaling methods, others (Devlin, et al., 1992) point out that such studies provide useful guidelines for practitioners and that more should be done. Continuing work needs to take place in applying IRT and other modern methodologies to the application of biographical information.

One additional point that deserves some attention is the fact that the analyses as they stand are limited by their very empirical nature. As pointed out earlier, the biographical instruments that were used were in no way tied to any sort of comprehensive job analysis of the position of air traffic controller. The BQ had been developed by the FAA, but a quick perusal of the items (Appendix A) defies their almost blatant exploratory nature. The original intent had been to locate items that demonstrated an

empirical relationship to future ATCS performance; and therefore may prove useful in future selection efforts. Though the ABA (Appendix B) had been developed around a construct-based framework (Gandy, Dye, & MacLane, 1994), it was intended for a fairly broad entry-level non-supervisory array of jobs. Prediction with a focus on the air traffic occupation was not part of the original game plan.

With this said, it is painfully obvious to this researcher that future efforts should be directed to developing an understanding of the job in question prior to any attempt to compare methods for scaling items. This understanding will include a thorough analysis of the job, with specifications for performance and what constitutes good and bad performance. Contextual factors for the job also present themselves as important to an understanding of what constitutes performance. Further, not only determining how performance is defined, but how it is measured will be paramount. Simultaneously, a thorough analysis of the "predictor space" should be conceptualized and operationalized, with a keen eye focused on the theoretical and empirical relationships between predictors and criteria.

Following these efforts, biodata item development can take place. This helps to ensure that the items will be grounded on some meaningful psychological foundation that truly can tap into elements of predictability. Pending this set of steps, a more meaningful and potentially accurate scaling comparison study can take place.

CHAPTER 6 – CONCLUSION

In conclusion, this study investigated some important sample characteristics of scored biodata keys on their ability to add incremental validity to an existing ability test for predicting school-based training performance. Further, this was extended by looking at the same for adding incremental validity to the ability test and school-based training performance for predicting on the job training (OJT) outcomes. Biodata key sample characteristics specifically looked at included: scaling method, key developmental sample size, and criterion characteristics.

Biodata surveys administered to air traffic control students at the FAA Academy between 1988 and 1990 were used to generate data that was scaled using three different measurement technologies: a) vertical percent empirical keying, b) factor analysis, and c) multidimensional item response theory (MIRT). Using the three methodologies, keys were created on developmental samples representing a variety of different *n*'s. These developmental sample sizes each represented approximately 66% of total sample sizes ranging from 4,568 to 100. The three resulting biodata scale scores were combined with an ability-based selection test to predict FAA Academy performance (defined as final score and pass/fail) and air traffic field facility OJT performance (defined as certification or "wash out"). Cross-validating the original developmental models on samples that represented approximately 33% of the aforementioned total samples, comparisons were made between each method, regarding the amount of shrinkage between developmental and cross-validation groups, magnitudes of resulting cross-validities, and classification accuracy.

The results indicated that at large sample sizes ($N \geq 1{,}000$) the differences between methods regarding the amount of shrinkage and the magnitude of cross-validity coefficients were negligible. At smaller sample sizes the amount of shrinkage increased with the empirical key enhanced model demonstrating a greater amount than the factor analytically derived keying model. Though differences were small, the incremental factor analytic model appeared to predict eventual performance, in most of the samples, better than the empirical or MIRT methods. This study contributes to the comparative scaling biodata literature and presents a practical application of multidimensional item response theory in a heretofore-undocumented area.

References

Aamodt, M.G., & Kimbrough, W.W. (1985). Comparison of four methods of weighting multiple predictors. *Educational and Psychological Measurement, 45,* 477-482.

Aamodt, M.G. & Pierce, W.L. (1987). Comparison of the rare response and vertical percent methods for scoring the biographical information blank. *Educational and Psychological Measurement, 47,* 505-511.

Ackerman, T.A. (1994). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement,* 18, 257-275.

Ackerman, T.A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7,* 255-278.

Ackerman, T.A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20,* 311-329.

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

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Asher, J.J. (1972). The biographical item: Can it be improved? *Personnel Psychology, 25,* 251-269.

Ashworth, S.D. (1989, April). The distinctions that Industrial/Organizational psychologists have between biodata and personality measurement are no longer meaningful. In T.W. Mitchell (chair). *Biodata vs. personality: The same or different classes of individual differences.* Symposium presented at the 4th annual conference of the Society for Industrial and Organizational Psychology, Boston MA.

Aul, J.C. (1998). Employing air traffic controllers, 1981-1992. In D. Broach (Ed.), *Recovery of the FAA air traffic control specialist workforce, 1981-1992* (DOT/FAA/AM 98/23) (pp. 3-6). Washington DC: FAA Office of Aviation Medicine.

Baehr, M.E., & Williams, G.B. (1967). Underlying dimensions of personal background data and their relationship to occupational classification. *Journal of Applied Psychology, 51,* 481-490.

Baehr, M.E., & Williams, G.B. (1968). Prediction of sales success from factorially determined dimensions of personal background data. *Journal of Applied Psychology, 52,* 98-103.

Barrick, M.R., & Mount, M.K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1-26.

Berkeley, M.H. (1952). *A comparison between the empirical and rational approaches for keying a heterogeneous test.* Doctoral dissertation, Washington University, St. Louis MO.

Bliesener, T. (1996). Methodological moderators in validating biographical data in personnel selection. *Journal of Occupational and Organizational Psychology, 69,* 107-120.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443-459.

Broach, D. (1993, Mar. 15). Personal communication.

Broach, D. (1998). Air traffic control specialist aptitude testing, 1981-1992. In D. Broach (Ed.), *Recovery of the FAA air traffic control specialist workforce, 1981-1992* (DOT/FAA/AM-98/23) (pp. 7-16). Washington DC: FAA Office of Aviation Medicine.

Broach, D., & Brecht-Clark, J. (1994). *Validation of the Federal Aviation Administration air traffic control specialist pre-training screen* (DOT/FAA/AM-94/4). Washington DC: FAA Office of Aviation Medicine.

Broach, D., Farmer, W.L., & Young, W.C. (1999). *Differential prediction of FAA Academy performance on the basis of race and written air traffic control specialist aptitude test scores* (DOT/FAA/AM-99/16). Washington DC: FAA Office of Aviation Medicine.

Broach, D., & Manning, C.A. (1997). *Review of air traffic controller selection: An International perspective* (DOT-FAA-AM-97-15). Washington DC: FAA Office Of Aviation Medicine.

Brokaw, L.D. (1959). *School and job validation of selection measures for air traffic control training* (WADC-TN-59-39). Lackland Air Force Base TX: Personnel Laboratory, Wright Air Development Center, USAF. (ASTIA No. AD 214 884)

Brown, D.C. (1994). Subgroup norming: Legitimate testing practice or reverse discrimination? *American Psychologist, 49*, 927-928.

Brown, S.H. (1994). Validating biodata. . In G.S.Stokes, M.D. Mumford, & W.A. Owens (Eds.), *Biodata handbook.* (pp. 199-236). Palo Alto CA: Consulting Psychologists Press, Inc.

Brush, D.H., & Owens, W.A. (1979). Implementation and evaluation of an assessment classification model for manpower utilization. *Personnel Psychology, 32*, 369-383.

Buss, A.H., & Plomin, R. (1975). *A temperament theory of personality development.* New York: Wiley.

Cascio, W.F. (1975). Accuracy of verifiable biographical information: Blank responses. *Journal of Applied Psychology, 60*, 767-769.

Cattell, R.B. (1943). The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology, 38*, 69-90.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

Cattell, R.B., & Coulter, M.A. (1966). Principles of behavioural taxonomy and the mathematical basis of the taxonome computer program. *British Journal of Mathematical and Statistical Psychology, 19*, 237-269.

Cattell, R.B., Coulter, M.A., & Tsujioka, B. (1966). The taxonometric recognition of types and functional emergents. In R.B. Cattell (Ed.). *Handbook of multivariate experimental psychology.* Chicago: Rand McNally.

Cattell, R.B., Eber, H.W., & Tatsuoka, M.M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire (16 PF).* Champaign IL: Institute for Personality and Ability Testing.

Chambers, J.A. (1964). Relating personality and biographical factors to scientific creativity. *Psychological Monographs, 78* (7, Whole No. 584).

Childs, A., & Klimoski, R.J. (1986). Successfully predicting career success: An application of the biographical inventory. *Journal of Applied Psychology, 71*, 3-8.

Clark, K.E. & Gee, H.H. (1954). Selecting keys for interest inventories. *Journal of Applied Psychology, 38*, 12-18.

Cleary, T.A. (1966). An individual differences model for multiple regression. *Psychometrika, 31,* 215-224.

Cobb, B.B., & Nelson, P.L. (1974). *Aircraft-pilot and other preemployment experience as factors in the selection of air traffic controller trainees* (FAA-AM-74-8). Washington DC: FAA Office of Aviation Medicine.

Cohen, J. & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale NJ: Lawrence Erlbaum Associates.

Collins, W.E., Boone, J.O., & VanDeventer, A.D. (Eds.) (1980). *The selection of air traffic control specialists: I. History and review of contributions by the Civil Aeromedical Institute* (DOT-FAA-AM-80-7). Washington DC: FAA Office of Aviation Medicine.(NTIS AD-A087655/7)

Collins, W.E., Boone, J.O., & VanDeventer, A.D. (1984). The selection of air traffic control specialists: Contributions by the Civil Aeromedical Institute. In S.B. Sells, J.T. Dailey, & E.W. Pickrel (Eds.), *Selection of air traffic controllers* (DOT-FAA-AM-84-2)(pp.79-111). Washington DC: FAA Office of Aviation Medicine. (NTIS AD-A147765)

Collins, W.E, Manning, C.A, & Taylor, D.K. (1984). A comparison of prestrike and Poststrike ATCS trainees: Biographic factors associated with Academy training success. In A.D.VanDeventer, W.E. Collins, C.A. Manning, D.K. Taylor, & N.E. Baxter, *Studies of poststrike air traffic control specialist trainees: I. Age, biographic factors, and selection test performance related to Academy training success* (FAA/AM-84-6) (pp. 7-14). Washington DC: FAA Office of Aviation Medicine.

Collins, W.E., Nye, L.G., & Manning, C.A. (1990). *Studies of poststrike air traffic control specialist trainees: III. Changes in demographic characteristics of Academy entrants and biodemographic predictors of success in air traffic controller selection and Academy screening* (DOT-FAA-AM-90-4). Washington DC: FAA Office of Aviation Medicine. (NTIS AD-A223480)

Collins, W.E., Schroeder, D.J., & Nye, L.G. (1989). *Relationships of anxiety scores to Academy and field training performance of air traffic control specialists* (DOT/FAA/AM-89-7). Washington DC: FAA Office of Aviation Medicine.

Coombs, C.H. (1964). *A theory of data.* New York: John Wiley & Sons.

Costa, P., Jr., & McCrae, R. (1985). *The NEO personality inventory (Form S).* Odessa FL: Psychological Assessment Resources.

Cowles, J.T., & Dailey, J.T. (1949, September). *The utility of biographical inventories in classification test batteries.* Paper presented at the annual meeting of the American Psychological Association, Denver CO.

Cronbach, L.J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671-684.

Dailey, C.A. (1960). The life history as a criterion of assessment. *Journal of Counseling Psychology, 7,* 20-23.

Dailey, J.T., & Pickrel, E.W. (1984). Development of the Multiplex Controller Aptitude Test. In S.B. Sells, J.T. Dailey, & E.W. Pickrel (Eds.), *Selection of air traffic controllers* (DOT-FAA-AM-84-2) (pp.281-297). Washington DC: FAA Office of Aviation Medicine. (NTIS AD-A147765)

Davis, K.R. (1984). A longitudinal analysis of biographical subgroups using Owens' developmental-integrative model. *Personnel Psychology, 37,* 1-14.

Dean, M.A. (1999). *On biodata construct validity, criterion-related validity, and adverse impact.* Unpublished doctoral dissertation. Louisiana State University, Baton Rouge.

Della Rocco, P.S. (1998). FAA Academy air traffic specialist screening programs and strike recovery. In D. Broach (Ed.), *Recovery of the FAA air traffic control specialist workforce, 1981-1992* (DOT/FAA/AM-98/23) (pp. 17-22). Washington DC: FAA Office of Aviation Medicine.

Della Rocco, P.S., Manning, C.A., & Wing, H. (1990). *Selection of air traffic controllers for automated systems: Applications from current research* (DOT/FAA/AM-90/13). Washington DC: FAA Office of Aviation Medicine.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39,* 1-38.

Devlin, S.E., Abrahams, N.M., & Edwards, J.E. (1992). Empirical keying of biographical data: Cross-validity as a function of scaling procedure and sample size. *Military Psychology, 4,* 119-136.

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

Digman, J.M., & Takemoto-Chock, N.K. (1981). Factors in the natural language of personality: Re-analysis and comparison of six major studies. *Multivariate Behavioral Research, 16,* 149-170.

Drasgow, F., & Hulin, C.L. (1990). Item response theory. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology: Vol. 1* (2nd ed., pp.577-636). Palo Alto CA: Consulting Psychologists Press, Inc.

DuBois, P.H., Loevinger, J., & Gleser, G.J. (1952). *The construction of homogeneous keys for a biographical inventory* (Research Bulletin 52-18). San Antonio TX: Personnel Research Laboratory, Lackland Air Force Base.

Dunnette, M.D. (1962). Personnel management. *Annual Review of Psychology, 13,* 285-314.

Dunnette, M.D. (1963). A modified model for test validation and selection research. *Journal of Applied Psychology, 47,* 317-323.

Eberhard, C., & Owens, W.A. (1975). Word association as a function of biodata subgrouping. *Developmental Psychology, 11,* 159-164.

Eberhardt, B.J., & Muchinsky, P.M. (1982a). An empirical investigation of the factor stability of Owens' Biographical Questionnaire. *Journal of Applied Psychology, 67,* 138-145.

Eberhardt, B.J., & Muchinsky, P.M. (1982b). Biodata determinants of vocational typology: An integration of two paradigms. *Journal of Applied Psychology, 67,* 714-727.

Eberhardt, B.J., & Muchinsky, P.M. (1984). Structural validation of Holland's hexagonal model: Vocational classification through the use of biodata. *Journal of Applied Psychology, 69,* 174-181.

England, G.W. (1971). *Development and use of weighted application blanks* (Rev. ed.). Minneapolis MN: Industrial Relations Center, University of Minnesota.

Equal Employment Opportunity Commission (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43,* 38290-38315.

Eysenck, H.J. (1944). Types of personality: A factorial study of seven-hundred neurotics. *Journal of Mental Science, 90,* 851-861.

Farmer, W.L., & Fiedler, E.R. (1999, August). *Personality and biodata overlap in air traffic controllers.* Presented at the 108th Annual Convention of the American Psychological Association, Boston MA.

Farmer, W.L., & Witt, L.A. (1998, April). User reactions to biodata, personality, and cognitive ability tests. In T.W. Mitchell (Chair), *The utility and practical value of biodata.* Symposium conducted at the 13th annual conference of the Society for Industrial and Organizational Psychology, Dallas TX.

Flanagan, J.C. (1947). Scientific development of the use of human resources: Progress in the Army Air Forces. *Science, 105,* 57-60.

Feild, H.S., Lissitz, R.W., & Schoenfeldt, L.F. (1975). The utility of homogeneous subgroups and individual information in prediction. *Multivariate Behavioral Research, 10,* 449-461.

Feild, H.S., & Schoenfeldt, L.F. (1975a). Development and application of a measure of students' college experiences. *Journal of Applied Psychology, 60,* 491-497.

Fine, S.A., & Cronshaw, S. (1994). The role of job analysis in establishing the validity of biodata. In G.S. Stokes, M.D. Mumford, & W.A. Owens (Eds.). *Biodata handbook.* (pp. 39-64). Palo Alto CA: Consulting Psychologists Press, Inc.

Fralicx, R.D. & Raju, N.S. (1982). A comparison of five methods for combining multiple criteria into a single composite. *Educational and Psychological Measurement, 42,* 823-827.

Fuentes, R.R., Sawyer, J.E., & Greener, J.M. (1989, August). *Comparison of the predictive characteristics of three biodata scaling methods.* Paper presented at the annual meeting of the American Psychological Association, New Orleans LA.

Fusilier, M.R., & Hoyer, W.D. (1980). Variables affecting perceptions of invasion of privacy in a personnel selection situation. *Journal of Applied Psychology, 65,* 623-626.

Gage, N.L. (1957). Logical versus empirical scoring keys: The case of the MTAI. *Journal of Educational Psychology, 48,* 213-216.

Gandy, J.A., Dye, D.A., & MacLane, C.N. (1994). Federal government selection: The individual achievement record. In G.S. Stokes, M.D. Mumford, & W.A. Owens (Eds.). *Biodata handbook.* (pp. 275-310). Palo Alto CA: Consulting Psychologists Press, Inc.

Ghiselli, E.E. (1956). Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology, 40,* 374-377.

Ghiselli, E.E. (1960a). Differentiation of tests in terms of the accuracy with which they predict for a given individual. *Educational and Psychological Measurement, 20,* 675-684.

Ghiselli, E.E. (1960b). The prediction of predictability. *Educational and Psychological Measurement, 20,* 3-8.

Ghiselli, E.E., Campbell, J.P., & Zedeck, S. (1981). *Measurement theory for the Behavioral sciences.* San Francisco: W.H. Freeman.

Glennon, J.R., Albright, L.E., & Owens, W.A. (1966). *A catalog of life history items.* Greensboro NC: Creativity Research Institute, Richardson Foundation.

Goldberg, L.R. (1972). Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monograph, 72-2.*

Goldsmith, D.B. (1922). The use of the personal history blank as a salesmanship test. *Journal of Applied Psychology, 6,* 149-155.

Gordon, R.A. (1997). Everyday life as an intelligence test: Effects of intelligence and intelligence context. *Intelligence, 24,* 203-320.

Gorsuch, R.L. (1983). *Factor analysis* (2nd ed.). Hillsdale NJ: Lawrence Erlbaum Associates.

Gottfredson, L.S. (1994). The science and politics of race-norming. *American Psychologist, 49,* 955-963.

Gottfredson, L.S. (1997). Why g matters: The complexity of everyday life. *Intelligence, 24,* 79-132.

Guilford, J.P. (1956). The structure of intellect. *Psychological Bulletin, 53,* 267-293.

Guilford, J.P. (1975). Factors and factors of personality. *Psychological Bulletin, 82,* 802-814.

Guion, R.M. (1965). *Personnel testing.* New York: McGraw-Hill.

Guion, R.M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah NJ: Lawrence Erlbaum Associates.

Gunter, B., Furnham, A., & Drakeley, R. (1993). *Biodata: Biographical indicators of business performance.* London: Routledge.

Hadley, J.M. (1944). The relation of personal data to achievement in a radio training school. *Psychological Bulletin, 41,* 60-63.

Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.)(pp. 147-200). New York: American Council on Education & Macmillan.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory.* Boston: Kluwer-Nijhoff.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park CA: Sage Publications.

Harman, H.H. (1976). *Modern factor analysis* (3rd ed. Rev.). Chicago: University of Chicago Press.

Harris, P.A. (1986). *A construct validity study of the Federal Aviation Administration Multiplex Controller Aptitude Test.* Washington DC: U.S. Office of Personnel Management.

Hase, H.D., & Goldberg, L.R. (1967). Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin, 67,* 231-248.

Hein, M., & Wesley, S. (1994). Scaling biodata through subgrouping. In G.S.Stokes, M.D. Mumford, & W.A. Owens (Eds.). *Biodata handbook.* (pp. 171-196). Palo Alto CA: Consulting Psychologists Press, Inc..

Henry, E.R. (1966). *Research conference on the use of autobiographical data as Psychological predictors.* Greensboro NC: Creativity Research Institute, Richardson Foundation.

Hilton, T.F., & Sells, S.B. (1984). Air traffic controller selection in the United States and other countries. An international overview. In S.B. Sells, J.T. Dailey, & E.W. Pickrel (Eds.). *Selection of air traffic controllers* (DOT-FAA-AM-84-2) (pp.26-37). Washington DC: FAA Office of Aviation Medicine. (NTIS AD-A147765)

Himmelstein, P., & Blaskovics, T.L. (1960). Prediction of an intermediate criterion of combat effectiveness with a biographical inventory. *Journal of Applied Psychology, 44,* 166-168.

Hogan, J.B. (1994). Empirical keying of background data measures. In G.S.Stokes, M.D. Mumford, & W.A. Owens (Eds.). *Biodata handbook.* (pp. 69-108). Palo Alto CA: Consulting Psychologists Press, Inc.

Hornick, C.W., James, L.R., & Jones, A.P. (1977). Empirical item keying versus a rational approach to analyzing a psychological climate questionnaire. *Applied Psychological Measurement, 1,* 489-500.

Hosmer, D.W., & Lemeshow, S. (1989). *Applied logistic regression.* New York: John Wiley & Sons.

Hough, L.M. (1984). Development and evaluation of the "accomplishment record" method of selecting and promoting professionals. *Journal of Applied Psychology, 69,* 135-146.

Hough, L., & Paullin, C. (1994). Construct-oriented scale construction: The rational approach. In G.S. Stokes, M.D. Mumford, & W.A. Owens (Eds.). *Biodata handbook.* (pp. 109-146). Palo Alto CA: Consulting Psychologists Press, Inc.

Hulin, C.L., Lissak, R.I., & Drasgow, F. (1982). Recovery of two- and three-parameter Logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6,* 249-260.

Jeanneret, P.R. (1997, Jul. 3). Personal communication.

Johnson, H.M. (1944). *On the actual and potential value of biographical information as a means of predicting success in aeronautical training* (CAA-RN-32). Washington DC: Airman Development Division, Civil Aeronautics Administration.

Keating, E., Paterson, D.G., & Stone, C.H. (1950). Validity of work histories obtained by interview. *Journal of Applied Psychology, 34,* 6-11.

Kelleher, E.J. (1972). Use of composite models in prediction. *Proceedings of the Annual Convention of the American Psychological Association, 7 (Pt.1),* 41-42.

Kelly, E.L., & Ewart, E. (1942). *A preliminary study of certain predictors of success in civilian pilot training* (CAA-RN-7). Washington DC: Division of Research, Civil Aeronautics Administration.

Kilcullen, R.N., White, L.A., Mumford, M.D., & Mack, H. (1995). Assessing the construct validity of rational biodata scales. *Military Psychology, 7,* 17-28.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis, 9,* 137-163.

Klein, S.P., & Owens, W.A., Jr. (1965). Faking of a scored life history blank as a function of criterion objectivity. *Journal of Applied Psychology, 49,* 452-454.

Klimoski, R.J. (1973). A biographical data analysis of career patterns in engineering. *Journal of Vocational Behavior, 3,* 103-113.

Korman, A.K. (1968). The prediction of managerial performance: A review. *Personnel Psychology, 21,*295-322.

Kuder, G.F. (1957). A comparative study of some methods of developing occupational keys. *Educational and Psychological Measurement, 17,* 105-114.

Kuhnert, K.W., & Russell, C.J. (1990). Using constructive developmental theory and biodata to bridge the gap between personnel selection and leadership. *Journal of Management, 16,* 595-607.

Lautenschlager, G.J., & Mendoza, J.L. (1986). A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. *Applied Psychological Measurement, 10,* 133-139.

Lautenschlager, G.J., & Shaffer, G.S. (1987). Reexamining the component stability of Owens' Biographical Questionnaire. *Journal of Applied Psychology, 72,* 149-152.

Lawshe, C.H., & Schucker, R.E. (1959). The relative efficiency of four test weighting methods in multiple prediction. Educational and Psychological Measurement, 19, 103-114.

Lecznar, W.B. (1951). *Evaluation of a new technique for keying biographical inventories empirically* (ATC HRRC Research Bulletin 51-2). Lackland Air Force Base TX: U.S. Air Force.

Lecznar, W.B., & Dailey, J.T. (1950). Keying biographical inventories in classification test batteries. *American Psychologist, 5,* 279.

Lecznar, W.B., Fructer, B., & Zachert, V. (1951). *A factor analysis of the Airman Biographical Inventory BE601B* (ATC HRRC Research Bulletin 51-3). Lackland Air Force Base TX: U.S. Air Force.

Ledvinka, J., & Scarpello, V.G. (1992). *Federal regulation of personnel and human Resource management* (3rd ed.). Belmont CA: Wadsworth.

Levine, A.S., & Zachert, V. (1951). Use of biographical inventory in the Air Force classification
program. *Journal of Applied Psychology, 35,* 241-244.

Lewin, K. (1936). *Principles of topological psychology.* New York: McGraw-Hill.

Lilienthal, M.G., & Pettyjohn, F.S. (1981). *Multiplex Controller Aptitude Test and Occupational Knowledge Test: Selection tools for air traffic controllers* (NAMRL Special Report 82-1). Pensacola FL: Naval Aerospace Medical Research Laboratory. (NTIS No.ADA118803)

Lissitz, R.W., & Schoenfeldt, L.F. (1974). Moderator subgroups for the estimation of educational performance: A comparison of prediction models. *American Educational Research Journal, 11*,63-75.

Loevinger, J., Gleser, G.C., & DuBois, P.H. (1953). Maximizing the discriminating power of a multiple-score test. *Psychometrika, 18,* 309-317.

Lockman, R.F. (1954). *Multivariate statistical analyses of naval aviation cadet selection measures* (Project Research Rep. No. NM 001 057.04.05). Pensacola FL: U.S. Naval School of Aviation Medicine.

Long, J.A., & Sandiford, P. (1935). *The validation of test items* (Research Bulletin No. 3). Toronto ON: University of Toronto, Department of Education.

Lord, F.M. (1952). A theory of test scores. *Psychometric Monographs*, No. 7.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

MacLane, C.N. (2000, Feb 17). Personal communication.

MacLane, C.N. (2000, Feb 22). Personal communication.

Mael, F.A. (1991). A conceptual rationale for the domain and attributes of biodata items. *Personnel Psychology, 44,* 763-792.

Mael, F.A., Connerly, M., & Morath, R.A. (1996). None of your business: Parameters of biodata invasiveness. *Personnel Psychology, 49,* 613-650.

Mael, F.A., & Schwartz, A.C. (1991). *Capturing temperament constructs with objective Biodata* (ARI Technical Report 939). Alexandria VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Magnusson, D. (1990). Personality development from an interactional perspective. In L.A. Pervin (Ed.), *Handbook of personality* (pp.193-222).

Malloy, J. (1955). The prediction of college achievement with the life experience inventory. *Educational and Psychological Measurement, 15,* 170-180.

Manning, C.A. (1991). Procedures for selection of air traffic control specialists. In H.Wing & C.A. Manning (Eds.), *Selection of air traffic controllers: complexity, requirements, and public interest* (DOT/FAA/AM-91/9) (pp. 13-22). Washington DC: FAA Office of Aviation Medicine.

Manning, C.A. (1998). Air traffic control specialist field training programs, 1981-1992. In D. Broach (Ed.), *Recovery of the FAA air traffic control specialist workforce, 1981-1992* (DOT/FAA/AM-98/23) (pp. 23-32). Washington DC: FAA Office of Aviation Medicine.

Manning, C.A., Della Rocco, P.S., & Bryant, K.D. (1989). *Prediction of success in FAA air traffic control field training as a function of selection and screening test performance* (DOT/FAA/AM-89/6). Washington DC: FAA Office of Aviation Medicine.

Manning, C.A., Kegg, P.S., & Collins, W.E. (1988). *Studies of poststrike air traffic control specialists: II. Selection and screening programs* (DOT/FAA/AM-88/3). Washington DC: FAA Office of Aviation Medicine.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Matteson, M.T. (1978). An alternative approach to using biographical data for predicting job success. *Journal of Occupational Psychology, 51,* 155-162.

McCrae, R.R., & Costa, P.T. (1985). Updating Norman's adequate taxonomy: Intelligence and personality dimensions in natural language and in questionnaires. *Journal of Personality and Social Psychology, 49,* 710-721.

McDaniel, M.A. (1989). Biographical constructs for predicting employee suitability. *Journal of Applied Psychology, 74,* 964-970.

McDonald, R.P. (1967). Nonlinear factor analysis. *Psychometric Monographs,* No. 15.

McDonald, R.P. (1968). A unified treatment of the weighting problem. *Psychometrika, 33,* 351-381.

McQuitty, L.L. (1957). Isolating predictor patterns associated with major criterion patterns. *Educational and Psychological Measurement, 17,* 3-42.

Mendoza, J.L. (2000, July 15). Personal communication.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education & Macmillan Publishing Company.

Mitchell, T.W. (1994). The utility of biodata. In G.S. Stokes, M.D. Mumford, & W.A. Owens (Eds.). *Biodata handbook.* (pp. 485-516). Palo Alto CA: Consulting Psychologists Press, Inc.

Mitchell, T.W. (1996). Can do and will do criterion success: A practitioner's theory of biodata. In R.B. Stennett, A.G. Parisi, & G.S. Stokes (Eds.), *A compendium: Papers presented to the first biennial biodata conference* (pp. 1-15). Athens GA: Applied Psychology Student Association, University of Georgia.

Mitchell, T.W. (chair) (1998, April). *The utility and practical value of biodata.* Symposium conducted at the 13th annual conference of the Society for Industrial and Organizational Psychology, Dallas TX.

Mitchell, T.W. (1998, Apr. 1). Personal communication.

Mitchell, T.W., & Klimoski, R.J. (1982). Is it rational to be empirical? A test of methods for scoring biographical data. *Journal of Applied Psychology, 67,* 411-418.

Mock, S.J. (1947). Biographical data. In, J.P. Guilford & J.I. Lacey (Eds:), *Printed classification tests. AAF Aviation Psychology Program Research Reports – Report No. 5.* Washington DC: U.S. Government Printing Office.

Morrison, R.F. (1977). A multivariate model for the occupational placement decision. *Journal of Applied Psychology, 62,* 271-277.

Morrison, R.F., Owens, W.A., Glennon, J.R., & Albright, L.E. (1962). Factored life history antecedents of industrial research performance. *Journal of Applied Psychology, 46,* 281-284.

Mosel, J.N., & Cozan, L.W. (1952). The accuracy of application blank work histories. *Journal of Applied Psychology, 36,* 365-369.

Mulaik, S.A. (1972). *A mathematical investigation of some multidimensional Rasch models for psychological tests.* Paper presented at the annual meeting of the Psychometric Society, Princeton NJ.

Mumford, M.D. (1999, Oct. 18). Personal communication.

Mumford, M.D., Costanza, D.P., Connelly, M.S., & Johnson, J.F. (1996). Item generation procedures and background data scales: Implications for construct and criterion related validity. *Personnel Psychology, 49,* 361-398.

Mumford, M.D., & Nickels, B.J. (1990). Making sense of people's lives: Applying principles of content and construct validity to background data. *Forensic Reports, 3,* 143-167.

Mumford, M.D., O'Connor, J., Clifton, T.C., Connelly, M.S., & Zaccaro, S.J. (1993). Background data constructs as predictors of leadership behavior. *Human Performance, 6,* 151-195.

Mumford, M.D., & Owens, W.A. (1984). Individuality in a developmental context: Some empirical and theoretical considerations. *Human Development, 27,* 84-108.

Mumford, M.D., & Owens, W.A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement, 11,* 1-31.

Mumford, M.D., Snell, A.F., & Reiter-Palmon, R. (1994). Personality and background data: Life history and self-concepts in an ecological system. In G.S. Stokes, M.D. Mumford, & W.A. Owens (Eds.). *Biodata handbook* (pp. 583-625). Palo Alto CA: Consulting Psychologists Press, Inc.

Mumford, M.D., & Stokes, G.S. (1992). Developmental determinants of individual action: Theory and practice in applying background measures. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology: Vol. 3* (2nd ed., pp.61-138). Palo Alto CA: Consulting Psychologists Press, Inc.

Mumford, M.D., Stokes, G.S., & Owens, W.A. (1990). *Patterns of life history: The ecology of human individuality.* Hillsdale NJ: Lawrence Erlbaum Associates.

Mumford, M.D., Stokes, G.S., Owens, W.A., & Jackson, K.E. (1990). Sequential results. In M.D. Mumford, G.S. Stokes, & W.A. Owens, *Patterns of life history: The ecology of human individuality* (pp. 149-193). Hillsdale NJ: Lawrence Erlbaum Associates.

Muraki, E. (1998, February 22). Personal communication.

Muraki, E., & Bock, R.D. (1997). *Parscale 3.* Chicago: Scientific Software International.

Muthen, B.O. (2001, June 16). Personal communication.

Muthen, L.K., & Muthen, B.O. (2001). *Mplus 2.* Los Angeles: Muthen & Muthen.

Nash, A.N. (1965). A study of item weights and scale lengths for the SVIB. *Journal of Applied Psychology, 49,* 264-269.

National Research Council (1946). *The history and development of the biographical inventory* (CAA-RN-70). Washington DC: Division of Research, Civil Aeronautics Administration.

Neidt, C.O., & Malloy, J.P. (1954). A technique for keying items of an inventory to be added to an existing test battery. *Journal of Applied Psychology, 38,* 308-312.

Neiner, A.G., & Owens, W.A. (1982). Relationships between two sets of biodata with 7 years separation. *Journal of Applied Psychology, 67,* 146-150.

Neiner, A.G., & Owens, W.A. (1985). Using biodata to predict job choice among college graduates. *Journal of Applied Psychology, 70,* 127-136.

Nickels, B.J. (1994) The nature of biodata. In G.S. Stokes, M.D. Mumford, & W.A. Owens (Eds.). *Biodata handbook.* (pp. 1-16). Palo Alto CA: Consulting Psychologists Press, Inc.

Norman, W.T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology, 66,* 574-583.

Norusis, M.J. (1994). *SPSS advanced statistics 7.* Chicago: SPSS.

Novick, M.R. (1974). Moderator subgroups and Bayesian *m*-group regression: Some concluding remarks. *American Educational Research Journal, 11,* 91-92.

Novick, M.R., & Jackson, P.H. (1974). Further cross-validation analysis of the Bayesian *m*-group regression method. *American Educational Research Journal, 11,* 77-85.

Nunnally, J.C. (1959). *Tests and measurement: Assessment and prediction.* New York: McGraw-Hill.

Nye, L.G., Schroeder, D.J., & Dollar, C.S. (1994). *Relationships of Type A behavior with biographical characteristics and training performances of air traffic controllers* (DOT-FAA-AM-94-13). Washington DC: FAA Office of Aviation Medicine. (NTIS AD-A283813)

O'Leary, L.R. (1973). Fair employment, sound psychometric practice, and reality. *American Psychologist, 28,* 147-150.

Osterlind, S.J. (1989). *Constructing test items: Multiple-choice, constructed response, performance, and other formats.* Boston: Kluwer.

Owens, W.A. (1953). Age and mental abilities: A longitudinal study. *Genetic Psychology Monographs, 48,* 3-54.

Owens, W.A. (1966). Age and mental abilities: A second adult follow-up. *Journal of Educational Psychology, 57,* 311-325.

Owens, W.A. (1968). Toward one discipline of scientific psychology. *American Psychologist, 23,* 782-785.

Owens, W.A. (1971). A quasi-actuarial basis for individual assessment. *American Psychologist,26,* 992-999.

Owens, W.A. (1976). Background data. In M.D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 609-644). Chicago: Rand McNally.

Owens, W.A. (1978). Moderators and subgroups. *Personnel Psychology, 31,* 243-247.

Owens, W.A., Glennon, J.R., & Albright, L.E. (1962). Retest consistency and the writing of life history items: A first step. *Journal of Applied Psychology, 46,* 329-331.

Owens, W.A., & Henry, E.R. (1966). *Biographical data in industrial psychology: A review and evaluation.* Greensboro NC: Creativity Research Institute, Richardson Foundation.

Owens, W.A., & Jewell, D.O. (1969). Personnel selection. *Annual Review of Psychology, 20,* 419-446.

Owens, W.A., & Schoenfeldt, L.F. (1979). Toward a classification of persons [Monograph]. *Journal of Applied Psychology, 64,* 569-607.

Owens, W.A., Schumacher, C.F., & Clark, J.B. (1957). The measurement of creativity in machine design. *Journal of Applied Psychology, 41,* 297-302.

Pace, L.A., & Schoenfeldt, L.F. (1977). Legal concerns in the use of weighted applications. *Personnel Psychology, 30,* 159-166.

Pervin, L.A. (1990). A brief history of modern personality theory. In L.A. Pervin (Ed.), *Handbook of Personality* (pp. 3-18). New York: The Guilford Press.

Pickrel, E.W. (1953). *The relative predictive efficiency of three methods of utilizing scores from biographical inventories.* Doctoral dissertation, University of Texas, Austin.

Reckase, M.D. (1972). *Development and application of a multivariate logistic latent trait model.* Unpublished doctoral dissertation, Syracuse University, Syracuse NY.

Reiter-Palmon, R. (1996). Background data factors revisited: The stability of Owens' biodata factors after 25 years. In R.B. Stennett, A.G. Parisi, & G.S. Stokes (Eds.), *A compendium: Papers presented to the first biennial biodata conference* (pp. 295-312). Athens GA: Applied Psychology Student Association, University of Georgia.

Rock, D.B., Dailey, J.T., Ozur, H., Boone, J.O., & Pickrel, E.W. (1981). *Selection of applicants for the air traffic controller occupation* (DOT-FAA-AM-82-11). Washington DC: FAA Office of Aviation Medicine. (NTIS AD-A122795/8)

Roskam, E.E. (1997). Models for speed and time-limit tests. In W.J. van der Linden & R.K. Hambleton (Eds.) (1997). Handbook of modern item response theory (pp. 187-208). New York: Springer-Verlag.

Rozeboom, W.W. (1979). Sensitivity of a linear composite of predictor items to differential item weighting. *Psychometrika, 44,* 289-296.

Russell, C.J. (1994). Generation procedures for biodata items: A point of departure. In G.S. Stokes, M.D. Mumford, & W.A. Owens (Eds.). *Biodata handbook.* (pp. 17-38). Palo Alto CA: Consulting Psychologists Press, Inc.

Russell, C.J. (1998, April 2). Personal communication.

Russell, C.J., Mattson, J., Devlin, S.E., & Atwater, D. (1990). Predictive validity of biodata items generated from retrospective life experience essays. *Journal of Applied Psychology, 75,* 569-580.

Russell, W., & Cope, G.V. (1925). A method of rating the history and achievements of applicants for positions. *Public Personnel Studies, 3,* 202-209.

Sackett, P.R., & Wilk, S.L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49,* 929-954.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometric Monographs,* No. 17.

Samejima, F. (1972). A general model for free-response data. *Psychometric Monographs*, No. 18.

Schmidt, F.L. (1988). The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behavior, 33*, 272-292.

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

Schneider, B. (1987a). The people make the place. *Personnel Psychology, 40*, 437-453.

Schneider, B. (1987b). E = f(P,B): The road to a radical approach to person-environment fit. *Journal of Vocational Behavior, 31*, 353-361.

Schneider, B., & Schneider, J.L. Biodata: An organizational focus. (1994). In G.S. Stokes, M.D. Mumford, & W.A. Owens (Eds.). *Biodata handbook.* (pp. 423-450). Palo Alto CA: Consulting Psychologists Press, Inc.

Schoenfeldt, L.F. (1974). Utilization of manpower: Development and evaluation of an assessment-classification model for matching individuals with jobs. *Journal of Applied Psychology, 59*, 583-595.

Schoenfeldt, L.F. (1996). From dust bowl empiricism to rational constructs in biodata. In R.B. Stennett, A.G. Parisi, & G.S. Stokes (Eds.), *A compendium: Papers presented to the first biennial biodata conference* (pp. 73-86). Athens GA: Applied Psychology Student Association, University of Georgia.

Schoenfeldt, L.F., & Lissitz, R.W. (1974). Moderator subgroups and Bayesian *m*-group regression: Some further comments. *American Educational Research Journal, 11*, 87-90.

Schoenfeldt, L.F., & Mendoza, J.L. (1994). Developing and using factorially derived biographical scales. In G.S. Stokes, M.D. Mumford, & W.A. Owens (Eds.). *Biodata handbook.* (pp. 147-170). Palo Alto CA: Consulting Psychologists Press, Inc.

Schooler, C. (1984). Psychological effects of complex environments during the life span: A review and theory. *Intelligence, 8*, 259-281.

Schroeder, D.J., Broach, D., & Young, W.C. (1993). *Contribution of personality to the prediction of success in initial air traffic control specialist training* (DOT-FAA AM-93-4). Washington DC: FAA Office of Aviation Medicine. (NTIS AD-A264699)

Schwab, D.P., & Oliver, R.L. (1974). Predicting tenure with biographical data: Exhuming buried evidence. *Personnel Psychology, 27,* 125-128.

Sells, S.B., & Pickrel, E.W. (1984). Introduction and overview. In S.B. Sells, J.T. Dailey, & E.W. Pickrel (Eds.), *Selection of air traffic controllers* (DOT-FAA-AM-84-2) (pp.1-23). Washington DC: FAA Office of Aviation Medicine. (NTIS AD-A147765)

Shaffer, G.S., Saunders, V., & Owens, W.A. (1986). Additional evidence for the accuracy of biographical data: Long-term retest and observer ratings. *Personnel Psychology, 39,* 791-809.

Shultz, K.S. (1996). Distinguishing personality and biodata items using confirmatory factor analysis of multitrait-multimethod matrices. *Journal of Business and Psychology, 10,* 263-288.

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

Stanley, J.C. & Wang, M.D. (1970). Weighting test items and test-item options, an overview of the analytical and empirical literature. *Educational and Psychological Measurement, 30,* 21-35.

Stead, W.H., Shartle, C.L., & Associates. (1940). *Occupational counseling techniques.* New York: American Book Company.

Steinhaus, S.D., & Waters, B.K. (1991). Biodata and the application of a psychometric perspective. *Military Psychology, 3,* 1-23.

Sternberg, R.J. (1985). Implicit theories of intelligence, creativity, and wisdom. *Journal of Personality and Social Psychology, 49,* 607-627.

Sternberg, R.J., & Grigorenko, E.L. (1997). Are cognitive styles still in style? *American Psychologist, 52,* 700-712.

Sternberg, R.J., & Lubart, T.I. (1996). Investing in creativity. *American Psychologist, 51,* 677-688.

Sternberg, R.J., Wagner, R.K., Williams, W.M., & Horvath, J.A. (1995). Testing common sense. *American Psychologist, 50,* 912-927.

Stokes, G.S. (1994). Introduction and history. In G.S. Stokes, M.D. Mumford, & W.A. Owens (Eds.). *Biodata handbook.* (pp. xv-xix). Palo Alto CA: Consulting Psychologists Press, Inc.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589-617.

Strimbu, J.L., & Schoenfeldt, L.F. (1973). Life history subgroups in the prediction of drug usage patterns and attitudes. *JSAS Catalog of Selected Documents in Psychology, 3* (MS. NO. 412), 83.

Strong, E.K., Jr (1926). An interest test for personnel managers. *Journal of Personnel Research, 5,* 194-204.

Sympson, J.B. (1978). A model for testing with multidimensional items. In D.J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82-98). Minneapolis: University of Minnesota, Department of Psychology.

Tabachnik, B.G., & Fidell, L.S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.

Taylor, D.K., VanDeventer, A.D., Collins, W.E., & Boone, J.O. (1983). Some biographical factors associated with success of air traffic control specialist trainees at the FAA Academy during 1980. In A.D. VanDeventer, D.K. Taylor, W.E. Collins, & J.O. Boone, (1983). *Three studies of biographical factors associated with success in air traffic control specialist screening/training at the FAA Academy* (DOT-FAA-AM-83-6) (pp. 6-11). Washington DC: FAA Office of Aviation Medicine. (NTIS AD-A128784/6)

Telenson, P.A., Alexander, R.A., & Barrett, G.V. (1983). Scoring the biographical information blank: A comparison of three weighting techniques. *Applied Psychological Measurement, 7,* 73-80.

Tenopyr, M.L. (1994). Big five, structural modeling, and item response theory. In G.S. Stokes, M.D. Mumford, & W.A. Owens (Eds.). *Biodata handbook.* (pp. 519-533). Palo Alto CA: Consulting Psychologists Press, Inc.

Terry, R.A. (1998, November 8). Personal communication.

Terry, R.A. (1999, October 18). Personal communication.

Tesser, A., & Lissitz, R.W. (1973). On an assumption underlying the use of homogeneous subgroups for prediction. *JSAS Catalog of Selected Documents in Psychology, 3* (MS. NO. 336), 38.

Thayer, P.W. (1977). "Somethings old, somethings new." *Personnel Psychology, 30,* 513-524.

Thissen, D. & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49,* 501-519.

Thurstone, L.L. (1938). Primary mental abilities. *Psychometric Monographs, No.1.*

Thurstone, L.L. (1947). *Multiple factor analysis.* Chicago: University of Chicago Press.

Toops, H.A. (1948). The use of addends in experimental control, social census, and managerial research. *Psychological Bulletin, 45,* 41-74.

Trattner, M.H. (1963). Comparison of three methods for assembling aptitude test batteries. *Personnel Psychology, 16,* 221-232.

Tupes, E.C., & Christal, R.C. (1961). *Recurrent personality factors based on trait ratings* (ASD-TR-61-97). Lackland Air Force Base TX: U.S. Air Force.

Tyler, L.E. (1959). Toward a workable psychology of individuality. *American Psychologist, 14,* 75-81.

van der Linden, W.J., & Hambleton, R.K. (Eds.) (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

VavDeventer, A.D. (1981, May). *Field training performance of FAA Academy air traffic control graduates.* Presented at the Annual Scientific Meeting of the Aerospace Medical Association.

VanDeventer, A.D. (1983a). Biographical profiles of successful and unsuccessful air traffic control specialist trainees. In A.D. VanDeventer, D.K. Taylor, W.E. Collins, & J.O. Boone, (1983). *Three studies of biographical factors associated with success in air traffic control specialist screening/training at the FAA Academy* (DOT-FAA-AM-83-6)(pp. 1-5). Washington DC: FAA Office of Aviation Medicine. (NTIS AD-A128784/6)

VanDeventer, A.D. (1983b). Military air traffic control experience and performance in FAA Academy air traffic control training. In A.D. VanDeventer, D.K. Taylor, W.E. Collins, & J.O. Boone, (1983). *Three studies of biographical factors associated with success in air traffic control specialist screening/training at the FAA Academy* (DOT-FAA-AM-83-6) (pp. 12-16). Washington DC: FAA Office of Aviation Medicine. (NTIS AD-A128784/6)

VanDeventer, A.D. (1984). A followup evaluation of the new aptitude testing procedures for selection of FAA air traffic control specialists. In A.D. VanDeventer, W.E. Collins, C.A.Manning, D.K. Taylor, & N.E. Baxter, *Studies of poststrike air traffic control specialist trainees: I. Age, biographic factors, and selection test performance related to Academy training success* (FAA/AM-84-6) (pp. 15-21). Washington DC: FAA Office of Aviation Medicine.

VanDeventer, A.D., & Baxter, N.E. (1984). Age and performance in air traffic control specialist training. In A.D. VanDeventer, W.E. Collins, C.A. Manning, D.K. Taylor, & N.E. Baxter, *Studies of poststrike air tarffic control specialist trainees: I. Age, biographic factors, and selection test performance related to Academy training success* (FAA/AM-84-6) (pp. 1-6). Washington DC: FAA Office of Aviation Medicine.

VanDeventer, A.D., Collins, W.E., Manning, C.A., Taylor, D.K., & Baxter, N.E. (1984). *Studies of poststrike air traffic control specialist trainees: I. Age, biographic factors, and selection test performance related to Academy training success* (DOT-FAA-AM-84-6). Washington DC: FAA Office of Aviation Medicine. (NTIS AD-A147892)

VanDeventer, A.D., Taylor, D.K., Collins, W.E., & Boone, J.O. (1983). *Three studies of biographical factors associated with success in air traffic control specialist screening/training at the FAA Academy* (DOT-FAA-AM-83-6). Washington DC: FAA Office of Aviation Medicine. (NTIS AD-A128784/6)

Verhelst, N.D., Verstralen, H.H.F.M., & Jansen, M.G.H. (1997). A logistic model for time limit tests. In W.J. van der Linden & R.K. Hambleton (Eds.) (1997). *Handbook of modern item response theory* (pp. 169-185). New York: Springer-Verlag.

Wagner, R.K., & Sternberg, R.J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49,* 436-458.

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

Wang, M.W., & Stanley, J.C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40,* 663-705.

Webb, S.C. (1960). The comparative validity of two biographical inventory keys. *Journal of Applied Psychology, 44,* 177-183.

Wernimont, P.F., & Campbell, J.P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52,* 372-376.

Whitely, S.E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,* 479-494.

Wilson D.T., Wood, R., & Gibbons, R. (1991). *Testfact.* Chicago: Scientific Software International.

Young, W.C., Broach, D., & Farmer, W.L. (1996). *Differential prediction of FAA Academy performance on the basis of gender and written air traffic control specialist aptitude test scores* (DOT/FAA/AM-96/13). Washington DC: FAA Office of Aviation Medicine.

Table 1

Breakdown of Selected Samples by "pass/fail" rates and developmental to cross-validation sample sizes

Dataset (sample *N*)	Pass (*n*/%)	Fail (*n*/%)	Sub-sample *n*'s Dev* (*n*) (~67%)	Sub-sample *n*'s CV* (*n*) (~33%)
Academy only	(~60%)	(~40%)		
4,568 (full)	2,732	1,836	3,114	1,454
3,000	1,800	1,200	1,983	1,017
2,000	1,199	801	1,345	655
1,500	923	577	992	508
1,000	594	406	663	337
500	307	193	340	160
100	61	39	63	37
Acad-Field	(~90%)	(~10%)		
2,000 (full)	1,815	185	1,329	671
1,500	1,356	144	997	503
1,000	912	88	688	312
500	456	44	318	182
100	90	10	67	33

* Dev = Developmental, CV = Cross-Validational

Table 2

Example weighting scheme for ABA item 20 (itemset 4,568 Academy only)

The high school math grade I most often received was:	*n* passed (%)	*n* failed	P% - F%	Net weight	Assigned weight
A	828 (45.12)	333 (26.14)	18.98	4	2
B	695 (37.97)	556 (43.56)	-5.59	-1	1
C	263 (14.29)	328 (25.67)	-11.37	-2	1
D or less	36 (1.96)	41 (3.14)	-1.18	0	1
Don't remember or Didn't take math	12 (.65)	20 (1.49)	-.84	-1	1

Table 3

Number of items remaining in empirically derived keys for each itemset

Dataset (sample *N*)	Number of items remaining
Academy only	
4,568 (full)	1
3,000	1
2,000	1
1,500	1
1,000	3
500	7
100	109
Acad-Field	
2,000 (full)	6
1,500	40
1,000	22
500	91
100	180

Table 4

Number of items remaining in each developmental sub-sample after deletion due to missing data and response invariance

Dataset (sample *N*)	Number of items remaining	% of full itemset (N_i = 296)
Academy only		
4,568 (full)	128	43.2
3,000	127	42.9
2,000	133	44.9
1,500	130	43.9
1,000	126	42.6
500	123	41.6
100	138	46.6
Acad-Field		
2,000 (full)	126	42.6
1,500	127	42.9
1,000	139	47.0
500	146	49.3
100	151	51.0

Table 5

Number of items remaining in each developmental sub-sample after deletion due to low inter-item correlation

Dataset (sample *N*)	Number of items remaining	% of full itemset (N_i = 296)
Academy only		
4,568 (full)	109	36.8
3,000	105	35.5
2,000	110	37.2
1,500	109	36.8
1,000	106	35.8
500	107	36.1
100	138	46.6
Acad-Field		
2,000 (full)	103	34.8
1,500	107	36.1
1,000	120	40.5
500	131	44.3
100	137	46.3

Table 6

Number of items remaining and components retained in each developmental sub-sample

Dataset (sample *N*)	Number of items remaining	Number of components	% of var. acctd. for	Mean inter-factor *r*
Academy only				
4,568 (full)	43	5	46.1	.11
3,000	41	5	46.9	.11
2,000	37	5	49.3	.12
1,500	41	5	48.4	.10
1,000	28	4	48.5	.09
500	25	4	46.8	.06
100	21	3	39.8	.06
Acad-Field				
2,000 (full)	40	6	50.5	.10
1,500	40	6	52.0	.09
1,000	37	6	53.3	.10
500	38	5	50.3	.12
100	32	4	45.9	.04

Table 7

Labels for and descriptions of retained components

Component label	Component description
GE – General Expectations	Reasons for seeking air traffic job. Expected job attributes, such as autonomy, decision-making potential, opportunity to work with others, etc.
WE – Work Ethic/Personal Characteristics	Self reported work ethics and personal attributes. Asked in the context of "how a former supervisor or other authority figure responsible for your performance would see you"
PE – Previous Air Traffic/Military Experience	Whether the individual had been in the military, and/or had any previous air traffic experience (most likely gained in the military) prior to the FAA.
HSAC – High School Academics	High school academic performance
HSAT – High School Athletics	Extent of high school athletic involvement
EA – Extracurricular Involvement	General involvement in extracurricular activities (e.g., clubs, teams, etc.)
ATT - Attendance	Self-reported attendance record
JS – Job Security Motivation	Job security (represented by benefit package, salary, etc.) as motivator of job choice
INT – Interpersonal Affiliation	Extent of social activities

Table 8

Retained components from each dataset

Dataset (sample *N*)	GE	WE	PE	HSAC	HSAT	EA	ATT	JS	INT
Academy only									
4,568 (full)	X (15, 12.8)	X (10, 10.3)	X (7, 8.7)	X (6, 8.2)	X (5, 6.2)				
3,000	X (13, 11.8)	X (10, 11.4)	X (7, 8.7)	X (6, 8.6)	X (5, 6.4)				
2,000	X (11, 12.2)	X (10, 12.0)	X (7, 9.8)	X (6, 9.4)	X (3, 5.8)				
1,500	X (13, 12.0)	X (11, 11.7)	X (7, 8.9)	X (7, 10.2)	X (3, 5.7)				
1,000	X (10, 16.7)	X (11, 14.1)	X (5, 11.7)	X (2, 9.6)					
500	X (10, 15.2)	X (7, 13.3)	X (6, 12.6)			X (2, 5.8)			
100	X (12, 19.6)	X (6, 12.1)	X (3, 8.1)						
Acad-Field									
2,000 (full)	X (12, 11.4)	X (10, 11.2)	X (8, 10.8)	X (5, 7.9)	X (3, 4.7)		X (2, 4.5)		
1,500	X (7, 10.7)	X (9, 10.1)	X (7, 9.6)	X (9, 10.0)	X (5, 6.5)			X (3, 5.1)	
1,000	X (11, 11.5)	X (9, 11.3)	X (6, 10.0)	X (6, 9.8)	X (3, 6.0)		X (2, 4.9)		
500	X (9, 12.1)	X (11, 11.7)	X (7, 11.5)	X (8, 11.1)	X (3, 8.0)				
100	X (10, 12.1)	X (12, 15.5)	X (7, 12.0)						X (3, 6.3)

Numbers in parentheses represent a) the number of items in each sample comprising the scale, and b) the percent of variance explained by each component, respectively.

Table 9

Number of items remaining in each developmental sub-sample after deletion due to low inter-item correlation ($r < .30$ criterion)

Dataset (sample *N*)	Number of items remaining	% of full itemset ($N_i = 296$)
Academy only		
4,568 (full)	111	37.5
3,000	109	36.8
2,000	111	37.5
1,500	110	37.2
1,000	108	36.5
500	108	36.5
100	140	47.3
Acad-Field		
2,000 (full)	106	35.8
1,500	108	36.5
1,000	124	41.9
500	132	44.6
100	138	46.6

Table 10

Number of items remaining in each developmental sub-sample after deletion due to low inter-item correlation ($r < .60$ criterion)

Dataset (sample *N*)	Number of items remaining	% of full itemset ($N_i = 296$)
Academy only		
4,568 (full)	90	30.4
3,000	88	29.7
2,000	90	30.4
1,500	99	33.4
1,000	85	28.7
500	61	20.6
100	42	14.1
Acad-Field		
2,000 (full)	92	31.1
1,500	92	31.1
1,000	94	31.8
500	78	26.7
100	38	12.8

Table 11

Number of items remaining and dimensions retained in each developmental sub-sample

Dataset (sample *N*)	Number of items remaining	Number of dimensions	% of var. acctd. for	Mean inter-factor *r*
Academy only				
4,568 (full)	27	5	64.4	.15
3,000	27	5	65.5	.16
2,000	27	5	63.9	.17
1,500	32	5	63.5	.15
1,000	*	*	*	*
500	*	*	*	*
100	*	*	*	*
Acad-Field				
2,000 (full)	29	5	64.0	.17
1,500	24	5	67.6	.13
1,000	*	*	*	*
500	*	*	*	*
100	*	*	*	*

* Convergence was not reached for this sample

Table 12

Retained components from each dataset

Dataset (sample *N*)	GE	WE	PE	HSAC	JS
Academy only					
4,568 (full)	X (7, 12.9)	X (5, 11.3)	X (6, 13.4)	X (6, 12.4)	X (3, 10.9)
3,000	X (6, 11.8)	X (5, 11.4)	X (7, 8.7)	X (7, 8.6)	X (2, 11.4)
2,000	X (7, 12.2)	X (5, 12.0)	X (6, 9.8)	X (6, 9.4)	X (3, 9.7)
1,500	X (8, 12.0)	X (6, 11.7)	X (7, 8.9)	X (7, 10.2)	X (4, 12.6)
1,000*					
500*					
100*					
Acad-Field					
2,000 (full)	X (12, 11.4)	X (10, 11.2)	X (8, 10.8)	X (5, 7.9)	X (2, 5.3)
1,500	X (7, 10.7)	X (9, 10.1)	X (7, 9.6)	X (9, 10.0)	X (3, 5.1)
1,000*					
500*					
100*					

Numbers in parentheses represent a) the number of items in each sample comprising the scale, and b) the percent of variancexplained by each component, respectively.
* No model convergence reached in these samples.

Table 13

Descriptive Characteristics for FAA Academy Sample (Original *N*=4,568)

Variable	Validation Sample	
	Developmental (*n*=3,114)	Cross-validation (*n*=1,454)
Academy Pass/Fail Status[1]	**1,836** (59.0) / **1,278** (41.0)	**896** (61.6) / **558** (38.4)
Academy Final Score[2]	**71.0 (11.5)**	**71.3 (11.4)**
	(29.9 - 95.9)	(30.1 – 95.7)
ATCS Entrance Exam[2]	**90.7 (4.8)**	**90.6 (4.6)**
	(70.6 - 100.0)	(73.8 – 100.0)
Biodata Score (empirical)[2]	**1.37 (.5)**	**1.36 (.5)**
	(1.0 - 2.0)	(1.0 – 2.0)
Factor Analytic Scores[2]		
General Expectations	**.01 (1.0)**	**-.02 (.9)**
	(-4.8 - 2.7)	(-3.4 – 2.9)
Work Ethic/Personal	**-.01 (1.0)**	**-.01 (1.0)**
Characteristics	(-4.3 - 2.3)	(-4.0 – 2.4)
Prior air traffic/Military	**-.01 (1.0)**	**.01 (1.0)**
Experience	(-2.1 - 3.1)	(-2.0 – 2.8)
High school academics	**.01 (1.0)**	**-.03 (1.0)**
	(-3.2 - 2.5)	(-3.1 – 2.5)
High school athletics	**-.01 (1.0)**	**.02 (.9)**
	(-2.3 - 2.7)	(-2.5 – 2.6)
MIRT Scores[2]		
High school academics	**-.01 (.7)**	**-.03 (.8)**
	(-2.4 - 1.6)	(-2.4 – 1.6)
Work ethic/personal	**-.03 (.8)**	**-.02 (.8)**
characteristics	(-2.8 - 1.8)	(-2.1 – 1.7)
General expectations	**-.04 (.7)**	**-.07 (.7)**
	(-2.7 - 1.5)	(-2.0 – 1.5)
Prior air traffic/Military	**.06 (.5)**	**.06 (.5)**
Experience	(-.6 - 1.3)	(-.6 – 1.2)
Job security importance	**-.05 (.6)**	**-.04 (.6)**
	(-1.9 - .9)	(-1.9 – 1.0)

[1] First number represents actual number, number in parentheses is percentage of sample
[2] First number is mean score, number in parentheses is standard deviation. Range below these numbers is score range.

Table 14

Descriptive Characteristics for FAA Academy Sample (Original N=3,000)

Variable	Validation Sample	
	Developmental (n=1,983)	Cross-validation (n=1,017)
Academy Pass/Fail Status[1]	**1,188** (59.9) / **795** (40.1)	**612** (60.2) / **405** (39.8)
Academy Final Score[2]	**71.2 (11.5)** (29.9 - 95.9)	**71.0 (11.5)** (32.6 – 95.7)
ATCS Entrance Exam[2]	**90.6 (4.7)** (73.2 - 100.0)	**90.5 (4.8)** (70.6 – 100.0)
Biodata Score (empirical)[2]	**1.36 (.5)** (1.0 - 2.0)	**1.36 (.5)** (1.0 – 2.0)
Factor Analytic Scores[2]		
General Expectations	**.00 (1.0)** (-4.7 - 2.5)	**-.02 (1.0)** (-3.7 – 2.7)
Work Ethic/Personal Characteristics	**-.01 (1.0)** (-3.3 - 2.2)	**.02 (1.0)** (-4.2 – 2.3)
Prior air traffic/Military Experience	**.01 (1.0)** (-1.7 – 2.8)	**-.02 (1.0)** (-1.7 – 2.8)
High school academics	**-.01 (1.0)** (-3.3 - 2.5)	**.03 (1.0)** (-3.4 – 2.1)
High school athletics	**.01 (1.0)** (-2.5 - 2.9)	**-.02 (1.0)** (-2.1 – 2.5)
MIRT Scores[2]		
High school academics	**-.03 (.7)** (-2.2 - 1.5)	**.02 (.7)** (-2.2 – 1.5)
Work ethic/personal characteristics	**-.01 (.7)** (-2.2 – 1.6)	**.01 (.8)** (-2.8 – 1.7)
General expectations	**-.03 (.7)** (-2.6 – 1.6)	**-.02 (.7)** (-1.9 – 1.6)
Prior air traffic/Military Experience	**.06 (.6)** (-.9 - 1.6)	**.03 (.6)** (-.9 – 1.6)
Job security importance	**-.05 (.6)** (-2.0 - .9)	**-.06 (.6)** (-1.9 – 0.9)

[1] First number represents actual number, number in parentheses is percentage of sample
[2] First number is mean score, number in parentheses is standard deviation. Range below these numbers is score range.

Table 15

Descriptive Characteristics for FAA Academy Sample (Original N=2,000)

Variable	Validation Sample	
	Developmental (n=1,345)	Cross-validation (n=655)
Academy Pass/Fail Status[1]	**819** (60.9) / **526** (39.1)	**380** (58.0) / **275** (42.0)
Academy Final Score[2]	**71.4 (11.7)**	**70.7 (11.1)**
	(29.9 - 95.8)	(35.6 – 92.8)
ATCS Entrance Exam[2]	**90.6 (4.6)**	**90.6 (4.7)**
	(72.2 – 100.0)	(73.5 – 100.0)
Biodata Score (empirical)[2]	**1.37 (.5)**	**1.35 (.5)**
	(1.0 - 2.0)	(1.0 – 2.0)
Factor Analytic Scores[2]		
General Expectations	**.00 (1.0)**	**-.01 (1.0)**
	(-4.1 - 2.8)	(-2.9 – 2.6)
Work Ethic/Personal	**.00 (1.0)**	**-.00 (1.0)**
Characteristics	(-3.8 - 2.3)	(-3.3 – 2.3)
Prior air traffic/Military	**-.02 (1.0)**	**.04 (1.0)**
Experience	(-2.1 - 3.3)	(-2.1 – 2.7)
High school academics	**.01 (1.0)**	**-.03 (1.0)**
	(-3.1 - 2.5)	(-2.6 – 2.5)
High school athletics	**.01 (1.0)**	**-.03 (1.0)**
	(-2.5 – 3.2)	(-2.3 – 2.6)
MIRT Scores[2]		
High school academics	**-.00 (.8)**	**-.04 (.8)**
	(-2.4 – 1.6)	(-2.4 – 1.6)
Work ethic/personal	**-.03 (.8)**	**-.03 (.8)**
characteristics	(-2.6 - 1.7)	(-2.4 – 1.7)
General expectations	**-.04 (.6)**	**-.02 (.6)**
	(-2.2 – 1.3)	(-1.7 – 1.3)
Prior air traffic/Military	**.03 (.4)**	**.02 (.4)**
Experience	(-.6 - 1.2)	(-.6 – 1.2)
Job security importance	**-.05 (.6)**	**-.06 (.6)**
	(-2.0 – 1.0)	(-1.7 – 1.0)

[1] First number represents actual number, number in parentheses is percentage of sample
[2] First number is mean score, number in parentheses is standard deviation. Range below these numbers is score range.

Table 16

Descriptive Characteristics for FAA Academy Sample (Original N=1,500)

Variable	Validation Sample: Developmental (n=992)	Validation Sample: Cross-validation (n=508)
Academy Pass/Fail Status[1]	**616** (62.1) / **376** (37.9)	**307** (60.4) / **201** (39.6)
Academy Final Score[2]	**71.7 (11.2)** (30.0 - 95.5)	**71.8 (11.3)** (35.6 – 95.8)
ATCS Entrance Exam[1]	**90.6 (4.8)** (70.6 - 100.0)	**90.9 (4.6)** (78.2 – 100.0)
Biodata Score (empirical)[2]	**1.39 (.5)** (1.0 - 2.0)	**1.36 (.5)** (1.0 – 2.0)
Factor Analytic Scores[2]		
General Expectations	**-.01 (1.0)** (-3.2 - 2.7)	**.01 (1.0)** (-3.5 – 2.7)
Work Ethic/Personal Characteristics	**.04 (1.0)** (-3.8 - 2.2)	**-.09 (1.0)** (-4.2 – 2.2)
Prior air traffic/Military Experience	**-.03 (1.0)** (-2.0 - 3.1)	**.06 (1.0)** (-2.0 – 2.5)
High school academics	**-.02 (1.0)** (-3.0 - 2.2)	**.05 (1.0)** (-3.2 – 2.2)
High school athletics	**-.02 (1.0)** (-2.4 – 3.1)	**.04 (.9)** (-2.2 – 2.3)
MIRT Scores[2]		
High school academics	**-.01 (.7)** (-2.2 - 1.5)	**-.01 (.7)** (-2.2 – 1.5)
Work ethic/personal characteristics	**-.08 (.7)** (-1.7 – 1.8)	**-.15 (.7)** (-2.7 – 1.6)
General expectations	**-.07 (.7)** (-2.0 – 1.5)	**-.05 (.6)** (-2.0 – 1.5)
Prior air traffic/Military Experience	**.03 (.4)** (-.7 - 1.3)	**.04 (.4)** (-.6 – 1.1)
Job security importance	**-.06 (.6)** (-2.0 – 1.0)	**-.08 (.6)** (-1.8 – 1.1)

[1] First number represents actual number, number in parentheses is percentage of sample
[2] First number is mean score, number in parentheses is standard deviation. Range below these numbers is score range.

Table 17

Descriptive Characteristics for FAA Academy Sample (Original *N*=1,000)

Variable	Validation Sample	
	Developmental (*n*=663)	Cross-validation (*n*=337)
Academy Pass/Fail Status[1]	**398** (60.0) / **265** (40.0)	**196** (58.2) / **141** (41.8)
Academy Final Score[2]	**70.7 (12.0)**	**71.0 (11.5)**
	(32.4 - 95.5)	(30.0 – 91.1)
ATCS Entrance Exam[2]	**90.9 (4.8)**	**90.5 (4.6)**
	(73.5 - 100.0)	(73.8 – 100.0)
Biodata Score (empirical)[2]	**3.79 (1.1)**	**3.80 (1.2)**
	(1.0 – 6.0)	(1.0 – 6.0)
Factor Analytic Scores[2]		
General Expectations	**-.01 (1.0)**	**.01 (.9)**
	(-3.9 - 2.6)	(-3.6 – 2.6)
Work Ethic/Personal	**.01 (1.0)**	**-.01 (.9)**
Characteristics	(-3.7 - 2.1)	(-2.9 – 2.1)
Prior air traffic/Military	**.00 (1.0)**	**-.01 (1.0)**
Experience	(-1.9 – 2.7)	(-1.8 – 2.5)
High school academics	**-.01 (1.0)**	**.03 (.9)**
	(-2.6 - 2.9)	(-2.3 – 2.7)

[1] First number represents actual number, number in parentheses is percentage of sample
[2] First number is mean score, number in parentheses is standard deviation. Range below these numbers is score range.

Table 18

Descriptive Characteristics for FAA Academy Sample (Original *N*=500)

Variable	Validation Sample	
	Developmental (*n*=340)	Cross-validation (*n*=160)
Academy Pass/Fail Status[1]	**217** (63.8) / **123** (36.2)	**90** (56.3) / **70** (43.8)
Academy Final Score[2]	**71.7 (11.5)**	**70.9 (12.1)**
	(36.3 - 95.3)	(35.0 – 91.4)
ATCS Entrance Exam[2]	**90.8 (4.8)**	**90.7 (4.8)**
	(75.1 - 100.0)	(73.2 – 100.0)
Biodata Score (empirical)[2]	**7.98 (1.7)**	**7.79 (1.5)**
	(3.0 - 12.0)	(4.0 – 12.0)
Factor Analytic Scores[2]		
General Expectations	**.03 (1.0)**	**-.07 (.9)**
	(-3.3 - 2.4)	(-2.7 – 2.2)
Work Ethic/Personal	**.02 (1.0)**	**-.03 (.9)**
Characteristics	(-3.1 - 2.1)	(-2.6 – 2.0)
Prior air traffic/Military	**-.02 (1.0)**	**.04 (1.0)**
Experience	(-2.5 – 2.6)	(-1.8 – 2.5)
Extracurricular Act.	**.07 (1.0)**	**-.14 (.9)**
	(-2.7 - 2.9)	(-2.4 – 2.0)

[1] First number represents actual number, number in parentheses is percentage of sample
[2] First number is mean score, number in parentheses is standard deviation. Range below these numbers is score range.

Table 19

Descriptive Characteristics for FAA Academy Sample (Original N=100)

Variable	Validation Sample	
	Developmental (n=63)	Cross-validation (n=37)
Academy Pass/Fail Status[1]	**39** (61.9) / **24** (38.1)	**22** (59.5) / **15** (40.5)
Academy Final Score[2]	**70.65 (10.8)** (43.0 – 91.1)	**73.55 (10.9)** (49.0 – 88.7)
ATCS Entrance Exam[2]	**90.83 (4.5)** (80.7 – 100.0)	**90.56 (5.1)** (79.2 – 100.0)
Biodata Score (empirical)[2]	**117.10 (16.6)** (77.0 – 146.0)	**117.68 (.5)** (75.0 – 134.0)
Factor Analytic Scores[2]		
General Expectations	**.01 (1.0)** (-1.8 – 1.9)	**-.03 (.8)** (-2.0 – 1.4)
Work Ethic/Personal Characteristics	**-.01 (1.0)** (-2.2 – 1.6)	**.04 (.8)** (-1.7 – 1.4)
Prior air traffic/Military Experience	**-.01 (1.0)** (-3.0 – 1.7)	**.09 (.8)** (-1.5 – 1.6)

[1] First number represents actual number, number in parentheses is percentage of sample
[2] First number is mean score, number in parentheses is standard deviation. Range below these numbers is score range.

Table 20

Intercorrelations of Relevant Variables for FAA Academy Sample (Original *N*=4,568)

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Academy		.51·	.36·	-.22·	.06·	-.14·	.30·	-.06·	.31·	.06·	-.24·	-.24·	-.16·
2. ATCS Exam	.20**		.14	-.07	.01	-.18	.14	-.02	.15	.00	-.09	-.20	-.06
3. Biodata	.23**	.14**		-.06	.08	-.07	.53	.00	.51	.12	-.06	-.21	-.06
4. FA-GE	-.16**	-.07**	-.06**		.00	-.00	.00	-.01	-.03	.13	.83	.02	.60
5. FA-WE	.06**	.01	.08**	.00		-.00	.00	.00	.19	.93	.16	.01	.01
6. FA-PE	.08**	-.18**	-.07**	-.00	-.00		.00	.00	-.12	.03	.04	.81	-.07
7. FA-HSAC	.17**	.14**	.53**	.00	.00	.00		-.00	.92	.08	-.08	-.34	-.11
8. FA-HSAT	-.04*	-.02	.00	-.01	.00	.00	-.00		.06	.05	.04	-.00	.10
9. MI-HSAC	.17**	.15**	.51**	-.03	.19**	-.12**	.92**	.06**		.26	-.04	-.40	-.05
10. MI-WE	.07**	.00	.12**	.13**	.93**	.03	.08**	.05**	.26**		.25	.02	.11
11. MI-GE	-.15**	-.09**	-.06**	.83**	.16**	.04	-.08**	.04**	-.04*	.25**		.08	.49
12. MI-PE	-.02	-.20**	-.21**	.02	.01	.81**	-.34**	-.00	-.40**	.02	.08**		.05
13. MI-JS	-.11**	-.06**	-.06**	.60**	.01	-.07**	-.11**	.10**	-.05**	.11**	.49**	.05**	

Sample n = 3,114; * $p \leq .05$ *(2-tailed)*, ** $p \leq .01$ *(2-tailed)*

Values in lower triangle are uncorrected; those in upper triangle are corrected for direct (*·) and indirect (*·) range restriction.

Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Table 21

Intercorrelations of Relevant Variables for FAA Academy Sample (Original *N*=3,000)

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Academy		.54·	.33·	-.18·	.11·	-.12·	.29·	-.09·	.29·	.12·	-.09·	-.20·	-.18·
2. ATCS Exam	.22**		.12	-.05	.03	-.15	.12	-.03	.14	.03	-.03	-.17	-.07
3. Biodata	.22**	.12**		-.06	.11	-.10	.55	.02	.51	.15	-.00	-.19	-.06
4. FA-GE	-.13**	-.05*	-.06**		.00	-.00	-.00	-.01	-.06	.10	.59	.09	.83
5. FA-WE	.08**	.03	.11**	.00		.00	.01	.00	.16	.94	.04	-.04	.15
6. FA-PE	.08**	-.15**	-.10**	-.00	.00		-.01	-.01	-.18	.02	-.05	.88	.11
7. FA-HSAC	.16**	.12**	.55**	-.00	.01	-.01		.00	.92	.07	-.02	-.23	-.04
8. FA-HSAT	-.06*	-.03	.02	-.01	.00	-.01	.00		.07	.00	-.01	-.01	-.01
9. MI-HSAC	.14**	.14**	.51**	-.06*	.16**	-.18**	.92**	.07**		.21	-.01	-.38	-.10
10. MI-WE	.09**	.03	.15**	.10**	.94**	.02	.07**	.00	.21**		.11	-.02	.22
11. MI-JS	-.05*	-.03	-.00	.59**	.04	-.05*	-.02	-.01	-.01	.11**		.05	.45
12. MI-PE	.01	-.17**	-.19**	.09**	-.04	.88**	-.23**	-.01	-.38**	-.02	.05*		.17
13. MI-GE	-.11**	-.07**	-.06*	.83**	.15**	.11**	-.04	-.01	-.10**	.22**	.45**	.17**	

Sample n = 1,983; * *p≤.05 (2-tailed),* ** *p≤.01 (2-tailed)*
Values in lower triangle are uncorrected; those in upper triangle are corrected for direct (*·) and indirect (*·) range restriction.
Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Table 22

Intercorrelations of Relevant Variables for FAA Academy Sample (Original *N*=2,000)

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Academy		.63·	.44·	-.23·	.07·	-.20·	.35·	-.02·	.33·	.10·	-.27·	-.18·	-.26·
2. ATCS Exam	.27**		.18	-.07	.01	-.17	.15	-.04	.15	.02	-.19	-.04	-.08
3. Biodata	.24**	.18**		-.03	.12	-.10	.57	-.01	.56	.18	-.23	-.02	-.06
4. FA-GE	-.15**	-.07*	-.03		.00	.01	.00	-.01	-.01	.09	-.02	.66	.83
5. FA-WE	.08**	.01	.12**	.00		.00	.01	.01	.19	.94	.01	.00	.13
6. FA-PE	.06*	-.17**	-.10**	.01	.00		-.00	.02	-.17	.02	.80	-.10	.01
7. FA-HSAC	.17**	.15**	.57**	.00	.01	-.00		.00	.91	.12	-.33	-.10	-.12
8. FA-HSAT	-.10**	-.04	-.01	-.01	.01	.02	.00		.08	.04	-.10	.01	.10
9. MI-HSAC	.14**	.15**	.56**	-.01	.19**	-.17**	.91**	.08**		.28	-.43	-.02	-.06
10. MI-WE	.09**	.02	.18**	.09**	.94**	.02	.12**	.04	.28**		-.01	.06	.17
11. MI-PE	-.01	-.19**	-.23**	-.02	.01	.80**	-.33**	-.10**	-.43**	-.01		-.01	.00
12. MI-JS	-.15**	-.04	-.02	.66**	.00	-.10**	-.10**	.01	-.02	.06	-.01		.56
13. MI-GE	-.18**	-.08**	-.06	.83**	.13**	.01	-.12**	.10**	-.06*	.17**	.00	.56**	

Sample n = 1,345; * $p \leq .05$ *(2-tailed),* ** $p \leq .01$ *(2-tailed)*

Values in lower triangle are uncorrected; those in upper triangle are corrected for direct (*·) and indirect (*·) range restriction.

Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Table 23

Intercorrelations of Relevant Variables for FAA Academy Sample (Original N=1,500)

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Academy		.59·	.39·	-.27·	-.03·	.31·	-.15·	-.12·	.34·	-.02·	-.24·	-.11·	-.31·
2. ATCS Exam	.25**		.16	-.10	-.03	.13	-.18	-.03	.16	-.03	-.19	-.02	-.13
3. Biodata	.23**	.16**		-.03	.06	.56	-.04	-.06	.58	.10	-.18	.00	-.04
4. FA-GE	-.17**	-.10**	-.03		-.00	.01	.01	-.01	-.05	.12	.10	.63	.81
5. FA-WE	.01	-.03	.06	-.00		.00	-.01	.00	.14	.93	.06	-.04	.17
6. FA-HSAC	.18**	.13**	.56**	.01	.00		.01	-.01	.94	.09	-.32	.02	.01
7. FA-PE	.11**	-.18**	-.04	.01	-.01	.01		-.02	-.15	.05	.78	-.08	.15
8. FA-HSAT	-.10**	-.03	-.06	-.01	.00	-.01	-.02		.05	.03	-.08	.01	.11
9. MI-HSAC	.16**	.16**	.58**	-.05	.14**	.94**	-.15**	.05		.20	-.42	.01	-.05
10. MI-WE	.03	-.03	.10**	.12**	.93**	.09**	.05	.03	.20**		.09	.04	.28
11. MI-PE	.01	-.19**	-.18**	.10**	.06	-.32**	.78**	-.08*	-.42**	.09**		.06	.18
12. MI-JS	-.10**	-.02	.00	.63**	-.04	.02	-.08*	.01	.01	.04	.06		.48
13. MI-GE	-.16**	-.13**	-.04	.81**	.17**	.01	.15**	.11**	-.05	.28**	.18**	.48**	

Sample n = 992; * *p≤.05 (2-tailed),* ** *p≤.01 (2-tailed)*

Values in lower triangle are uncorrected; those in upper triangle are corrected for direct (*·) and indirect (*·) range restriction.

Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Table 24

Intercorrelations of Relevant Variables for FAA Academy Sample (Original N=1,000)

	1	2	3	4	5	6	7
1. Academy		.44·	.39·	-.17·	.20·	-.20·	.03·
2. ATCS Exam	.17**		.15	-.03	.08	-.23	.01
3. Biodata	.28**	.15**		-.12	.23	-.26	.29
4. FA-GE	-.15**	-.03	-.12**		.01	-.00	.00
5. FA-WE	.13**	.08*	.23**	.01		.00	-.01
6. FA-PE	.01	-.23**	-.26**	-.00	.00		.02
7. FA-HSAC	.03	.01	.29**	.00	-.01	.02	

Sample n = 663; * $p \leq .05$ *(2-tailed)*, ** $p \leq .01$ *(2-tailed)*
Values in lower triangle are uncorrected; those in upper triangle are corrected for direct (*·) and indirect (*) range restriction.
Biodata = Empirical Key, FA = Factor Analytic Key.
FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics.

Table 25

Intercorrelations of Relevant Variables for FAA Academy Sample (Original *N*=500)

	1	2	3	4	5	6	7
1. Academy		.50·	.36·	-.34·	.11·	-.21·	.13·
2. ATCS Exam	.20**		.07	-.11	.00	-.25	.06
3. Biodata	.31**	.07		-.35	.44	-.05	.11
4. FA-GE	-.24**	-.11*	-.35**		.01	-.00	-.03
5. FA-WE	.13*	.00	.44**	.01		.01	.01
6. FA-PE	.06	-.25**	-.05	-.00	.01		-.02
7. FA-EA	.07	.06	.11*	-.03	.01	-.02	

Sample n = 340; * $p \leq .05$ *(2-tailed)*, ** $p \leq .01$ *(2-tailed)*
Values in lower triangle are uncorrected; those in upper triangle are corrected for direct (*·) and indirect (*·) range restriction.
Biodata = Empirical Key, FA = Factor Analytic Key.
FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, EA=Extracurricular Activity Involvement.

Table 26

Intercorrelations of Relevant Variables for FAA Academy Sample (Original N=100)

	1	2	3	4	5	6
1. Academy		.81·	.84·	-.37·	.01·	.56·
2. ATCS Exam	.43**		.32	-.10	.02	.32
3. Biodata	.69**	.32**		-.56	-.07	.23
4. FA-GE	-.28*	-.10	-.56**		-.05	.01
5. FA-WE	-.06	.02	-.07	-.05		-.00
6. FA-PE	.12	.32*	.23	.01	-.00	

Sample n = 63; * $p \leq .05$ *(2-tailed)*, ** $p \leq .01$ *(2-tailed)*
Values in lower triangle are uncorrected; those in upper triangle are corrected for direct (*·) and indirect (*·) range restriction.
Biodata = Empirical Key, FA = Factor Analytic Key.
FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience.

Table 27

Hierarchical Linear Regression Analysis for prediction of FAA Academy performance from ATCS Entrance Exam and Biodata Keys – Original N = 4,568 (Uncorrected)

Step and Variable	β	t	R^2	ΔR^2	ΔF
Step 1					
ATCS Entrance Exam	.204	11.21***	.042	.042	125.73***
Step 2a (Empirical)					
ATCS Entrance Exam	.174	9.68***	.083	.041	131.52***
Biodata Key	.207	11.47***			
Step 2b (FA)					
ATCS Entrance Exam	.193	10.63***	.102	.060	39.09***
FA-GE	-.145	-8.23***			
FA-WE	.062	3.53***			
FA-PE	.111	6.18***			
FA-HSAC	.147	8.26***			
FA-HSAT	-.041	-2.34***			
Step 2c (MIRT)					
ATCS Entrance Exam	.184	10.11***	.090	.048	31.04***
MI-HSAC	.153	7.50***			
MI-WE	.064	3.32***			
MI-GE	-.141	-6.67***			
MI-PE	.085	4.31***			
MI-JS	-.028	-1.36			

Sample n = 3,114; *$p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$*
Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Table 28

Hierarchical Linear Regression Analysis for prediction of FAA Academy performance from ATCS Entrance Exam and Biodata Keys – Original N = 3,000 (Uncorrected)

Step and Variable	β	t	R^2	ΔR^2	ΔF
Step 1					
ATCS Entrance Exam	.216	9.46***	.047	.047	89.46***
Step 2a (Empirical)					
ATCS Entrance Exam	.193	8.57***	.086	.039	77.72***
Biodata Key	.199	8.82***			
Step 2b (FA)					
ATCS Entrance Exam	.206	9.08***	.100	.053	21.49***
FA-GE	-.116	-5.21***			
FA-WE	.069	3.12**			
FA-PE	.112	4.97***			
FA-HSAC	.139	6.20***			
FA-HSAT	-.055	-2.48*			
Step 2c (MIRT)					
ATCS Entrance Exam	.207	9.10***	.087	.040	16.10***
MI-HSAC	.127	5.10***			
MI-WE	.090	3.80***			
MI-JS	-.006	-0.23			
MI-PE	.117	4.74***			
MI-GE	-.126	-4.84***			

Sample n = 1,983; *$p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$*
Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Table 29

Hierarchical Linear Regression Analysis for prediction of FAA Academy performance from ATCS Entrance Exam and Biodata Keys – Original N = 2,000 (Uncorrected)

Step and Variable	β	t	R^2	ΔR^2	ΔF
Step 1					
ATCS Entrance Exam	.268	9.86***	.072	.072	97.16***
Step 2a (Empirical)					
ATCS Entrance Exam	.232	8.54***	.110	.108	53.33***
Biodata Key	.198	7.30***			
Step 2b (FA)					
ATCS Entrance Exam	.253	9.31***	.131	.059	17.01***
FA-GE	-.136	-5.14***			
FA-WE	.078	2.96**			
FA-PE	.105	3.91***			
FA-HSAC	.126	4.72***			
FA-HSAT	-.095	-3.61***			
Step 2c (MIRT)					
ATCS Entrance Exam	.256	9.41***	.123	.051	14.38***
MI-HSAC	.118	3.75***			
MI-WE	.073	2.55*			
MI-PE	.096	3.19***			
MI-JS	-.071	-2.20*			
MI-GE	-.121	-3.70***			

Sample n = 1,345; *$p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$*
Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Table 30

Hierarchical Linear Regression Analysis for prediction of FAA Academy performance from ATCS Entrance Exam and Biodata Keys – Original N = 1,500 (Uncorrected)

Step and Variable	β	t	R^2	ΔR^2	ΔF
Step 1					
ATCS Entrance Exam	.252	7.86***	.064	.064	61.74***
Step 2a (Empirical)					
ATCS Entrance Exam	.221	6.93***	.100	.037	37.06***
Biodata Key	.194	6.09***			
Step 2b (FA)					
ATCS Entrance Exam	.245	7.70***	.139	.076	15.84***
FA-GE	-.153	-4.93***			
FA-WE	.017	0.54			
FA-HSAC	.147	4.71***			
FA-PE	.149	4.75***			
FA-HSAT	-.091	-2.93**			
Step 2c (MIRT)					
ATCS Entrance Exam	.234	7.26***	.115	.052	10.50***
MI-HSAC	.179	4.95***			
MI-WE	.023	0.68			
MI-PE	.152	4.27***			
MI-JS	-.035	-0.96			
MI-GE	-.137	-3.59***			

Sample n = 992; *$p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$*
Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Table 31

Hierarchical Linear Regression Analysis for prediction of FAA Academy performance from ATCS Entrance Exam and Biodata Keys – Original N = 1,000 (Uncorrected)

Step and Variable	β	t	R^2	ΔR^2	ΔF
Step 1					
ATCS Entrance Exam	.171	4.34***	.029	.029	18.82***
Step 2a (Empirical)					
ATCS Entrance Exam	.131	3.41***	.096	.067	45.84***
Biodata Key	.261	6.77***			
Step 2b (FA)					
ATCS Entrance Exam	.167	4.19***	.067	.038	6.26***
FA-GE	-.147	-3.78***			
FA-WE	.116	2.99**			
FA-PE	.045	1.13			
FA-HSAC	.028	0.71			

Sample n = 663; $p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$
Biodata = Empirical Key, FA = Factor Analytic Key.
FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics.

Table 32

Hierarchical Linear Regression Analysis for prediction of FAA Academy performance from ATCS Entrance Exam and Biodata Keys – Original $N = 500$ (Uncorrected)

Step and Variable	β	t	R^2	ΔR^2	ΔF
Step 1					
ATCS Entrance Exam	.201	3.62***	.040	.040	13.07***
Step 2a (Empirical)					
ATCS Entrance Exam	.180	3.39***	.130	.090	32.05***
Biodata Key	.300	5.66***			
Step 2b (FA)					
ATCS Entrance Exam	.200	3.59***	.121	.081	7.10***
FA-GE	-.221	-4.11***			
FA-WE	.129	2.41*			
FA-PE	.110	1.99*			
FA-EA	.058	1.08			

Sample $n = 340$; $p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$
Biodata = Empirical Key, FA = Factor Analytic Key.
FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, EA=Extracurricular Activity Involvement.

Table 33

Hierarchical Linear Regression Analysis for prediction of FAA Academy performance from ATCS Entrance Exam and Biodata Keys – Original $N = 100$ (Uncorrected)

Step and Variable	β	t	R^2	ΔR^2	ΔF
Step 1					
ATCS Entrance Exam	.428	3.54***	.183	.183	12.56***
Step 2a (Empirical)					
ATCS Entrance Exam	.233	2.36*	.521	.337	38.70***
Biodata Key	.613	6.22***			
Step 2b (FA)					
ATCS Entrance Exam	.409	3.24**	.247	.063	1.49
FA-GE	-.242	-2.01*			
FA-WE	-.085	-0.71			
FA-PE	-.006	-0.05			

Sample $n = 63$; $p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$
Biodata = Empirical Key, FA = Factor Analytic Key.
FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience.

Table 34

Hierarchical Linear Regression Analysis for prediction of FAA Academy performance from ATCS Entrance Exam and Biodata Keys – Original $N = 4{,}568$ (Corrected)

Step and Variable	β	t	R^2	ΔR^2	ΔF
Step 1					
ATCS Entrance Exam	.505	31.48***	.255	.255	990.69***
Step 2a (Empirical)					
ATCS Entrance Exam	.464	30.28***	.337	.082	358.01***
Biodata Key	.290	18.92***			
Step 2b (FA)					
ATCS Entrance Exam	.448	29.15***	.355	.100	89.57***
FA-GE	-.188	-12.56***			
FA-WE	.059	3.94***			
FA-PE	-.057	-3.77***			
FA-HSAC	.242	16.01***			
FA-HSAT	-.054	-3.60***			
Step 2c (MIRT)					
ATCS Entrance Exam	.448	29.12***	.352	.097	85.80***
MI-HSAC	.195	11.32***			
MI-WE	.059	3.62***			
MI-GE	-.179	-10.04***			
MI-PE	-.054	-3.25***			
MI-JS	-.042	-2.42*			

Sample $n = 3{,}114$; *$p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$*
Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Table 35

Hierarchical Linear Regression Analysis for prediction of FAA Academy performance from ATCS Entrance Exam and Biodata Keys – Original N = 3,000 (Corrected)

Step and Variable	β	t	R^2	ΔR^2	ΔF
Step 1					
ATCS Entrance Exam	.536	27.15***	.287	.287	736.88***
Step 2a (Empirical)					
ATCS Entrance Exam	.505	26.78***	.361	.073	208.73***
Biodata Key	.272	14.45***			
Step 2b (FA)					
ATCS Entrance Exam	.489	25.91***	.377	.090	52.81***
FA-GE	-.151	-8.18***			
FA-WE	.092	4.98***			
FA-PE	-.046	-2.49**			
FA-HSAC	.230	12.33***			
FA-HSAT	-.077	-4.19***			
Step 2c (MIRT)					
ATCS Entrance Exam	.495	25.94***	.360	.075	42.58***
MI-HSAC	.183	8.77***			
MI-WE	.100	5.07***			
MI-JS	-.018	-0.88			
MI-PE	-.016	-0.77			
MI-GE	-.144	-6.58***			

Sample n = 1,983; *$p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$*
Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Table 36

Hierarchical Linear Regression Analysis for prediction of FAA Academy performance from ATCS Entrance Exam and Biodata Keys – Original N = 2,000 (Corrected)

Step and Variable	β	t	R^2	ΔR^2	ΔF
Step 1					
ATCS Entrance Exam	.628	28.56***	.394	.394	815.97***
Step 2a (Empirical)					
ATCS Entrance Exam	.567	27.92***	.501	.108	270.55***
Biodata Key	.334	16.45***			
Step 2b (FA)					
ATCS Entrance Exam	.557	27.31***	.511	.117	59.50***
FA-GE	-.189	-9.53***			
FA-WE	.073	3.66***			
FA-PE	-.108	-5.37***			
FA-HSAC	.262	13.09***			
FA-HSAT	.009	0.44			
Step 2c (MIRT)					
ATCS Entrance Exam	.556	27.67***	.504	.110	55.20***
MI-HSAC	.170	7.21***			
MI-WE	.079	3.66***			
MI-PE	-.091	-4.00***			
MI-JS	-.048	-1.98*			
MI-GE	-.195	-7.91***			

Sample n = 1,345; *$p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$*
Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Table 37

Hierarchical Linear Regression Analysis for prediction of FAA Academy performance from ATCS Entrance Exam and Biodata Keys – Original N = 1,500 (Corrected)

Step and Variable	β	t	R^2	ΔR^2	ΔF
Step 1					
ATCS Entrance Exam	.590	22.02***	.348	.348	484.85***
Step 2a (Empirical)					
ATCS Entrance Exam	.541	21.47***	.438	.090	144.67***
Biodata Key	.303	12.03***			
Step 2b (FA)					
ATCS Entrance Exam	.525	21.03***	.469	.121	41.01***
FA-GE	-.221	-9.07***			
FA-WE	-.019	-0.78			
FA-HSAC	.248	10.15***			
FA-PE	-.059	-2.41**			
FA-HSAT	-.107	-4.41***			
Step 2c (MIRT)					
ATCS Entrance Exam	.520	20.62***	.460	.112	37.42***
MI-HSAC	.244	8.65***			
MI-WE	.007	0.27			
MI-PE	.000	0.02			
MI-JS	.009	0.33			
MI-GE	-.232	-7.80***			

Sample n = 992; *p≤.05 (2-tailed), ** p≤.01 (2-tailed), ***p≤.001*
Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Table 38

Hierarchical Linear Regression Analysis for prediction of FAA Academy performance from ATCS Entrance Exam and Biodata Keys – Original N = 1,000 (Corrected)

Step and Variable	β	t	R^2	ΔR^2	ΔF
Step 1					
ATCS Entrance Exam	.439	12.21***	.193	.193	148.97***
Step 2a (Empirical)					
ATCS Entrance Exam	.389	11.46***	.300	.108	95.85***
Biodata Key	.332	9.79***			
Step 2b (FA)					
ATCS Entrance Exam	.395	11.07***	.258	.065	13.68***
FA-GE	-.157	-4.52***			
FA-WE	.171	4.94***			
FA-PE	-.115	-3.23***			
FA-HSAC	.035	1.01			

Sample n = 663; *$p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$*
Biodata = Empirical Key, FA = Factor Analytic Key.
FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics.

Table 39

Hierarchical Linear Regression Analysis for prediction of FAA Academy performance from ATCS Entrance Exam and Biodata Keys – Original $N = 500$ (Corrected)

Step and Variable	β	t	R^2	ΔR^2	ΔF
Step 1					
ATCS Entrance Exam	.497	10.12***	.247	.247	102.35***
Step 2a (Empirical)					
ATCS Entrance Exam	.475	10.36***	.350	.103	49.47***
Biodata Key	.322	7.03***			
Step 2b (FA)					
ATCS Entrance Exam	.435	9.16***	.360	.113	13.53***
FA-GE	-.288	-6.28***			
FA-WE	.117	2.56*			
FA-PE	-.098	-2.08*			
FA-EA	.097	2.12*			

Sample $n = 340$; $p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$
Biodata = Empirical Key, FA = Factor Analytic Key.
FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, EA=Extracurricular Activity Involvement.

Table 40

Hierarchical Linear Regression Analysis for prediction of FAA Academy performance from ATCS Entrance Exam and Biodata Keys – Original N = 100 (Corrected)

Step and Variable	β	t	R^2	ΔR^2	ΔF
Step 1					
ATCS Entrance Exam	.813	10.45***	.661	.661	109.18***
Step 2a (Empirical)					
ATCS Entrance Exam	.607		1.000	.339	
Biodata Key	.644				
Step 2b (FA)					
ATCS Entrance Exam	.673	12.27***	.858	.197	24.57***
FA-GE	-.307	-5.90***			
FA-WE	-.019	-0.364			
FA-PE	.352	6.45***			

Sample n = 63; $p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$
Biodata = Empirical Key, FA = Factor Analytic Key.
FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience.

Table 41

Hierarchical Logistic Regression Analysis for prediction of FAA Academy pass/fail status from ATCS Entrance Exam and Biodata Keys – Original $N = 4{,}568$

Step and Variable	e^B	Wald χ^2	-2LL	Δ-2LL	[1]Model R^2
Step 1					
ATCS Entrance Exam	1.07	71.26***	3845.97	[2]74.14***	.034
Step 2a (Empirical)					
ATCS Entrance Exam	1.06	50.78***	3743.96	102.61***	.075
Biodata Key	2.20	89.49***			
Step 2b (FA)					
ATCS Entrance Exam	1.07	62.36***	3742.87	103.10***	.080
FA-GE	.79	35.52***			
FA-WE	1.11	6.56**			
FA-PE	1.18	17.51***			
FA-HSAC	1.29	41.09***			
FA-HSAT	.97	.54			
Step 2c (MIRT)					
ATCS Entrance Exam	1.07	56.17***	3754.70	91.27***	.071
MI-HSAC	1.43	34.86***			
MI-WE	1.14	5.63*			
MI-GE	.72	24.74***			
MI-PE	1.31	7.99**			
MI-JS	.98	.05			

Sample $n = 3{,}114$; *$p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$*

Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Notes: (1) Model R^2 as defined in Nagelkerke (1991); (2) Δ-2LL for Step 1 represents the difference between the –2LL for the model with no independent variables and that for the model with one (ATCS Entrance Exam) independent variable.

Table 42

Hierarchical Logistic Regression Analysis for prediction of FAA Academy pass/fail status from ATCS Entrance Exam and Biodata Keys – Original N = 3,000

Step and Variable	e^B	Wald χ^2	-2LL	Δ-2LL	[1]Model R^2
Step 1					
ATCS Entrance Exam	1.08	55.45***	2408.97	[2]58.25***	.042
Step 2a (Empirical)					
ATCS Entrance Exam	1.08	44.48***	2348.00	60.97***	.081
Biodata Key	2.17	52.82***			
Step 2b (FA)					
ATCS Entrance Exam	1.08	49.57***	2347.68	61.29***	.085
FA-GE	.78	20.36***			
FA-WE	1.15	8.25**			
FA-PE	1.17	9.36**			
FA-HSAC	1.23	18.07***			
FA-HSAT	.90	4.63*			
Step 2c (MIRT)					
ATCS Entrance Exam	1.08	49.47***	2363.28	45.69***	.072
MI-HSAC	1.26	9.72**			
MI-WE	1.27	11.57***			
MI-JS	.95	.23			
MI-PE	1.28	6.42*			
MI-GE	.74	13.06***			

Sample n = 1,983; $p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$

Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Notes: (1) Model R^2 as defined in Nagelkerke (1991); (2) Δ-2LL for Step 1 represents the difference between the –2LL for the model with no independent variables and that for the model with one (ATCS Entrance Exam) independent variable.

Table 43

Hierarchical Logistic Regression Analysis for prediction of FAA Academy pass/fail status from ATCS Entrance Exam and Biodata Keys – Original $N = 2{,}000$

Step and Variable	e^B	Wald χ^2	-2LL	Δ-2LL	[1]Model R^2
Step 1					
ATCS Entrance Exam	1.10	49.51***	1627.06	[2]52.87***	.056
Step 2a (Empirical)					
ATCS Entrance Exam	1.09	35.43***	1586.19	40.87***	.098
Biodata Key	2.27	39.25***			
Step 2b (FA)					
ATCS Entrance Exam	1.10	42.42***	1574.60	52.46***	.109
FA-GE	.78	16.70***			
FA-WE	1.25	13.55***			
FA-PE	1.14	4.31*			
FA-HSAC	1.22	10.22**			
FA-HSAT	.84	8.62**			
Step 2c (MIRT)					
ATCS Entrance Exam	1.10	44.08***	1577.68	49.38***	.099
MI-HSAC	1.27	6.94**			
MI-WE	1.29	8.81**			
MI-PE	1.36	3.28			
MI-JS	.79	3.61			
MI-GE	.76	5.25*			

Sample $n = 1{,}345$; *$p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$*
Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics. JS=Job Security as motivator of job choice.
Notes: (1) Model R^2 as defined in Nagelkerke (1991); (2) *Δ-2LL* for Step 1 represents the difference between the *–2LL* for the model with no independent variables and that for the model with one (ATCS Entrance Exam) independent variable.

Table 44

Hierarchical Logistic Regression Analysis for prediction of FAA Academy pass/fail status from ATCS Entrance Exam and Biodata Keys – Original N = 1,500

Step and Variable	e^B	Wald χ^2	-2LL	Δ-2LL	[1]Model R^2
Step 1					
ATCS Entrance Exam	1.10	35.68***	1167.75	[2]38.07***	.056
Step 2a (Empirical)					
ATCS Entrance Exam	1.09	28.13***	1146.53	21.22***	.086
Biodata Key	1.98	20.62***			
Step 2b (FA)					
ATCS Entrance Exam	1.10	34.10***	1130.76	36.99***	.108
FA-GE	.79	11.02**			
FA-WE	1.12	2.31			
FA-HSAC	1.26	10.44**			
FA-PE	1.25	9.06**			
FA-HSAT	.86	4.22*			
Step 2c (MIRT)					
ATCS Entrance Exam	1.09	30.94***	1140.12	27.63***	.094
MI-HSAC	1.39	8.15**			
MI-WE	1.21	2.88			
MI-PE	1.60	5.05*			
MI-JS	.82	2.21			
MI-GE	.74	5.30*			

Sample n = 992; $p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$

Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Notes: (1) Model R^2 as defined in Nagelkerke (1991); (2) Δ-2LL for Step 1 represents the difference between the –2LL for the model with no independent variables and that for the model with one (ATCS Entrance Exam) independent variable.

Table 45

Hierarchical Logistic Regression Analysis for prediction of FAA Academy pass/fail status from ATCS Entrance Exam and Biodata Keys – Original N = 1,000

Step and Variable	e^B	Wald χ^2	-2LL	Δ-2LL	[1]Model R^2
Step 1					
ATCS Entrance Exam	1.06	9.92**	832.15	[2]10.14**	.022
Step 2a (Empirical)					
ATCS Entrance Exam	1.04	5.30*	798.89	33.26***	.091
Biodata Key	1.58	29.46***			
Step 2b (FA)					
ATCS Entrance Exam	1.06	10.23**	805.49	26.66***	.077
FA-GE	.69	18.26***			
FA-WE	1.21	5.03*			
FA-PE	1.14	2.27			
FA-HSAC	1.04	.26			

Sample n = 663; *$p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$*
Biodata = Empirical Key, FA = Factor Analytic Key.
FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics.
Notes: (1) Model R^2 as defined in Nagelkerke (1991); (2) Δ-2LL for Step 1 represents the difference between the –2LL for the model with no independent variables and that for the model with one (ATCS Entrance Exam) independent variable.

Table 46

Hierarchical Logistic Regression Analysis for prediction of FAA Academy pass/fail status from ATCS Entrance Exam and Biodata Keys – Original $N = 500$

Step and Variable	e^B	Wald χ^2	-2LL	Δ-2LL	[1]Model R^2
Step 1					
ATCS Entrance Exam	1.10	12.92***	399.84	[2]13.80***	.059
Step 2a (Empirical)					
ATCS Entrance Exam	1.10	11.46**	352.43	47.41***	.242
Biodata Key	1.75	38.75***			
Step 2b (FA)					
ATCS Entrance Exam	1.10	11.32**	376.44	23.40***	.153
FA-GE	.69	8.62**			
FA-WE	1.53	11.72**			
FA-PE	1.09	.44			
FA-EA	1.16	1.45			

Sample $n = 340$; *p≤.05 (2-tailed), ** p≤.01 (2-tailed), ***p≤.001*
Biodata = Empirical Key, FA = Factor Analytic Key.
FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, EA=Extracurricular Activity Involvement.
Notes: (1) Model R^2 as defined in Nagelkerke (1991); (2) Δ-2LL for Step 1 represents the difference between the –2LL for the model with no independent variables and that for the model with one (ATCS Entrance Exam) independent variable.

Table 47

Hierarchical Logistic Regression Analysis for prediction of FAA Academy pass/fail status from ATCS Entrance Exam and Biodata Keys – Original $N = 100$

Step and Variable	e^B	Wald χ^2	-2LL	Δ-2LL	[1]Model R^2
Step 1					
ATCS Entrance Exam	1.21	6.32*	70.09	[2]7.81**	.171
Step 2a (Empirical)					
ATCS Entrance Exam	2.69	2.61	6.64	63.45***	.957
Biodata Key	2.35	3.25			
Step 2b (FA)					
ATCS Entrance Exam	1.20	5.25*	65.76	4.33	.26
FA-GE	.53	3.58			
FA-WE	.80	.60			
FA-PE	1.09	.09			

Sample $n = 63$; *$p \leq .05$ (2-tailed), **$p \leq .01$ (2-tailed), ***$p \leq .001$
Biodata = Empirical Key, FA = Factor Analytic Key.
FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience.
Notes: (1) Model R^2 as defined in Nagelkerke (1991); (2) Δ-2LL for Step 1 represents the difference between the –2LL for the model with no independent variables and that for the model with one (ATCS Entrance Exam) independent variable.

Table 48

FAA Academy (Score) Validities and Cross-Validities as a Function of Keying Method and Sample Size (Uncorrected)

	Samples													
	4568		3000		2000		1500		1000		500		100	
	V	CV	V	CV	V	CV	V	CV	V	CV	V	CV	V	CV
Sample Size	3114	1454	1983	1017	1345	655	992	508	663	337	340	160	63	37
ATCS Exam	.20	.26	.22	.21	.27	.22	.25	.22	.17	.23	.20	.12	.43	.10
Exam + Biodata														
Empirical	.29	.31	.29	.29	.33	.28	.32	.30	.31	.30	**.36**[1]	**.04**[1]	**.72**[1]	**.25**[1]
FA	.32	.32	.32	.33	.36	.33	.37	.35	.26	.27	**.35**[1]	**.10**[1]	**.50**[2]	**.08**[2]
MIRT	.30	.30	.30	.30	.35	.30	.34	.32	*[3]	*[3]	*[3]	*[3]	*[3]	*[3]

FAA Academy final score used as criterion. Linear regression weights applied.

Note: V = validity; CV = cross-validity. (1) Difference between validity and cross-validity is significant at $p \leq .01$; (2) Difference between validity and cross-validity is significant at $p \leq .05$; (3) No keys using the MIRT were available at this sample size.

Table 49

FAA Academy (Score) Cross-Validities as a Function of Keying Method and Sample Size (Uncorrected)

	Samples						
	4568	3000	2000	1500	1000	500	100
Sample Size	1454	1017	655	508	337	160	37
ATCS Exam	.26**	.21**	.22**	.22**	.23**	.12	.10
Exam + Biodata							
Empirical	.31**	.29**	.28**	.30**	.30**	.04	.25
FA	.32**	.33**	.33**	.35**	.27**	.10	.08
MIRT	.30**	.30**	.30**	.32**	*[1]	*[1]	*[1]

** $p \leq .01$

FAA Academy final score used as criterion. Linear regression weights applied.

Note: (1) No keys using the MIRT were available at this sample size.

Table 50

FAA Academy (Score) Validities and Cross-Validities as a Function of Keying Method and Sample Size (Corrected)

	Samples													
	4568		3000		2000		1500		1000		500		100	
	V	CV	V	CV	V	CV	V	CV	V	CV	V	CV	V	CV
Sample Size	3114	1454	1983	1017	1345	655	992	508	663	337	340	160	63	37
ATCS Exam	.20	.26	.22	.21	.27	.22	.25	.22	.17	.23	.20	.12	.43	.10
Exam + Biodata														
Empirical	.28	.32	.28	.27	.33	.28	.31	.29	.29	.31	**.33**[1]	**.08**[1]	**.69**[1]	**.20**[1]
FA	.28	.29	.28	.27	.31	.26	.32	.30	.22	.24	.29	.12	.45	.16
MIRT	.27	.27	.27	.26	.32	.25	.31	.30	*[2]	*[2]	*[2]	*[2]	*[2]	*[2]

FAA Academy final score used as criterion. Linear regression weights applied.

Note: V = validity; CV = cross-validity. (1) Difference between validity and cross-validity is significant at $p \leq .01$; (2) No keys using the MIRT were available at this sample size.

Table 51

FAA Academy (Score) Cross-Validities as a Function of Keying Method and Sample Size (Corrected)

	Samples						
	4568	3000	2000	1500	1000	500	100
Sample Size	1454	1017	655	508	337	160	37
ATCS Exam	.26**	.21**	.22**	.22**	.23**	.12	.10
Exam + Biodata							
Empirical	.32**	.27**	.28**	.29**	.31**	.08	.20
FA	.29**	.27**	.26**	.30**	.24**	.12	.16
MIRT	.27**	.26**	.25**	.30**	*[1]	*[1]	*[1]

** $p \leq .01$

FAA Academy final score used as criterion. Linear regression weights applied.
Note: (1) No keys using the MIRT were available at this sample size.

Table 52

FAA Academy (Pass/Fail) Validities and Cross-Validities as a Function of Keying Method and Sample Size

	Samples													
	4568		3000		2000		1500		1000		500		100	
	V	CV	V	CV	V	CV	V	CV	V	CV	V	CV	V	CV
Sample Size	3114	1454	1983	1017	1345	655	992	508	663	337	340	160	63	37
ATCS Exam	.16	.20	.18	.15	.20	.19	.20	.17	.13	.18	.20	.10	.34	.14
Exam + Biodata														
Empirical	.23	.23	.24	.20	.27	.25	.25	.24	.25	.25	**.43**[1]	**.04**[1]	**.96**[1]	**.15**[1]
FA	.24	.24	.26	.23	.28	.28	.28	.27	.25	.19	**.34**[1]	**.11**[1]	.42	.13
MIRT	.23	.22	.23	.22	.27	.26	.26	.24	*[2]	*[2]	*[2]	*[2]	*[2]	*[2]

FAA Academy pass/fail status used as criterion. Logistic regression weights applied.

Note: V = validity; CV = cross-validity. (1) Difference between validity and cross-validity is significant at $p \leq .01$; (2) No keys using the MIRT were available at this sample size.

Table 53

FAA Academy (Pass/Fail Status) Cross-Validities as a Function of Keying Method and Sample Size

	Samples						
	4568	3000	2000	1500	1000	500	100
Sample Size	1454	1017	655	508	337	160	37
ATCS Exam	.20**	.15**	.19**	.17**	.18**	.10	.14
Exam + Biodata							
Empirical	.23**	.20**	.25**	.24**	.25**	.04	.15
FA	.24**	.23**	.28**	.27**	.19**	.11	.13
MIRT	.22**	.22**	.26**	.24**	*[1]	*[1]	*[1]

** $p \leq .01$

FAA Academy pass/fail status used as criterion. Logistic regression weights applied.

Note: (1) No keys using the MIRT were available at this sample size.

Table 54

FAA Academy (Pass/Fail Status) Classification Accuracy as a Function of Keying Method and Sample Size

	Samples						
	4568	3000	2000	1500	1000	500	100
Sample Size	1454	1017	655	508	337	160	37
ATCS Exam	61.1	59.2	58.5	59.8	58.1	55.9	64.7
Exam + Biodata							
Empirical	63.4	61.8	62.7	62.0	58.1	55.9	61.8
FA	62.9	62.1	64.4	61.5	58.5	62.1	58.8
MIRT	62.1	61.1	64.1	61.7	*[1]	*[1]	*[1]

FAA Academy pass/fail status used as criterion. Logistic regression weights applied to individual exam and biodata scores to calculate predicted probability of passing. Number represents percentage of individuals correctly classified.
Note: (1) No keys using the MIRT were available at this sample size.

Table 55

Descriptive Characteristics for FAA Field Facility Sample (Original *N*=2,000)

Variable	Validation Sample	
	Developmental (*n*=1,329)	Cross-validation (*n*=671)
ATCS Cert./Wash Out[1]	**1,204** (90.6) / **125** (9.5)	**611** (91.1) / **60** (8.9)
Academy Final Score[2]	**79.1 (5.6)**	**78.9 (6.0)**
	(70.0 - 95.2)	(70.0 – 96.0)
ATCS Entrance Exam[2]	**91.1 (4.8)**	**91.4 (4.8)**
	(73.2 - 100.0)	(75.1 – 100.0)
Biodata Score (empirical)[2]	**5.26 (1.2)**	**5.24 (1.2)**
	(1.0 – 8.0)	(1.0 – 8.0)
Factor Analytic Scores[2]		
General Expectations	**.02 (1.0)**	**-.03 (1.0)**
	(-4.0 - 2.8)	(-4.2 – 2.3)
Work Ethic/Personal	**.02 (1.0)**	**-.04 (1.0)**
Characteristics	(-3.5 - 2.2)	(-3.6 – 2.3)
Prior air traffic/Military	**.02 (1.0)**	**-.03 (.9)**
Experience	(-1.7 – 2.8)	(-1.6 – 2.5)
High school academics	**.02 (1.0)**	**-.04 (1.0)**
	(-2.8 - 2.2)	(-3.1 – 2.1)
High school athletics	**-.02 (1.0)**	**.04 (1.0)**
	(-2.4 – 3.5)	(-2.3 – 3.9)
Attendance	**.01 (1.0)**	**-.01 (1.0)**
	(-4.9 – 2.9)	(-4.3 – 2.6)
MIRT Scores[2]		
High school academics	**-.03 (.8)**	**-.06 (.8)**
	(-2.2 – 1.6)	(-2.4 – 1.5)
Work ethic/personal	**-.05 (.7)**	**-.11 (.8)**
characteristics	(-2.2 – 1.6)	(-2.8 – 1.6)
General expectations	**-.08 (.7)**	**-.09 (.7)**
	(-2.2 – 1.6)	(-2.4 – 1.5)
Prior air traffic/Military	**.09 (.7)**	**.03 (.6)**
Experience	(-1.1 - 1.8)	(-1.2 – 1.8)
Job security importance	**-.06 (.7)**	**-.09 (.7)**
	(-2.6 – 1.3)	(-2.4 – 1.3)

[1] First number represents actual number, number in parentheses is percentage of sample
[2] First number is mean score, number in parentheses is standard deviation. Range below these numbers is score range.

Table 56

Descriptive Characteristics for FAA Field Facility Sample (Original *N*=1,500)

Variable	Validation Sample	
	Developmental (*n*=997)	Cross-validation (*n*=503)
ATCS Cert./Wash Out[1]	**903** (90.6) / **94** (9.4)	**453** (90.1) / **50** (9.9)
Academy Final Score[2]	**79.1 (5.8)**	**79.1 (5.7)**
	(70.0 – 96.0)	(70.0 – 94.7)
ATCS Entrance Exam[2]	**91.1 (4.8)**	**91.2 (4.7)**
	(73.2 - 100.0)	(76.6 – 100.0)
Biodata Score (empirical)[2]	**44.37 (11.3)**	**44.29 (11.2)**
	(20.0 – 74.0)	(21.0 – 76.0)
Factor Analytic Scores[2]		
General Expectations	**.00 (1.0)**	**-.00 (1.0)**
	(-4.5 - 2.5)	(-3.6 – 2.6)
Work Ethic/Personal	**.02 (1.0)**	**-.04 (1.0)**
Characteristics	(-3.7 - 2.4)	(-3.6 – 2.0)
Prior air traffic/Military	**-.00 (1.0)**	**.01 (.9)**
Experience	(-1.7 – 2.8)	(-1.5 – 2.5)
High school academics	**-.03 (1.0)**	**.06 (1.0)**
	(-2.7 - 2.2)	(-2.9 – 2.1)
High school athletics	**-.02 (1.0)**	**.04 (1.0)**
	(-2.5 – 3.0)	(-2.2 – 2.7)
Job security importance	**-.01 (1.0)**	**-.01 (1.0)**
	(-3.6 – 2.7)	(-3.3 – 2.5)
MIRT Scores[2]		
High school academics	**-.03 (.7)**	**.04 (.7)**
	(-2.1 – 1.6)	(-2.0 – 1.5)
Work ethic/personal	**-.05 (.8)**	**-.06 (.8)**
characteristics	(-2.3 – 1.4)	(-2.9 – 1.4)
General expectations	**-.02 (.7)**	**-.02 (.7)**
	(-2.3 – 1.4)	(-2.3 – 1.4)
Prior air traffic/Military	**.05 (.5)**	**.03 (.5)**
Experience	(-.7 - 1.3)	(-.7 – 1.3)
Job security importance	**-.02 (.5)**	**.02 (.6)**
	(-1.8 – .9)	(-1.8 – .9)

[1] First number represents actual number, number in parentheses is percentage of sample
[2] First number is mean score, number in parentheses is standard deviation. Range below these numbers is score range.

Table 57

Descriptive Characteristics for FAA Field Facility Sample (Original *N*=1,000)

Variable	Validation Sample	
	Developmental (*n*=688)	Cross-validation (*n*=312)
ATCS Cert./Wash Out[1]	**627** (91.1) / **61** (8.9)	**285** (91.3) / **27** (8.7)
Academy Final Score[2]	**78.9 (5.7)**	**78.8 (5.9)**
	(70.1 - 95.5)	(70.1 – 95.7)
ATCS Entrance Exam[2]	**91.3 (4.8)**	**90.8 (5.1)**
	(74.4 - 100.0)	(73.2 – 100.0)
Biodata Score (empirical)[2]	**22.49 (3.4)**	**22.41 (3.3)**
	(10.0 – 31.0)	(11.0 – 31.0)
Factor Analytic Scores[2]		
General Expectations	**-.00 (1.0)**	**.01 (1.0)**
	(-3.3 - 2.9)	(-4.0 – 2.7)
Work Ethic/Personal	**-.01 (1.0)**	**.02 (1.0)**
Characteristics	(-3.5 - 2.2)	(-3.5 – 2.0)
Prior air traffic/Military	**.01 (1.0)**	**-.02 (.9)**
Experience	(-1.7 – 2.8)	(-1.7 – 2.3)
High school academics	**.02 (1.0)**	**-.03 (1.0)**
	(-3.2 - 2.1)	(-3.0 – 1.9)
High school athletics	**.01 (1.0)**	**-.03 (.9)**
	(-2.3 – 4.2)	(-2.0 – 3.5)
Attendance	**.01 (1.0)**	**-.03 (1.0)**
	(-3.8 – 2.5)	(-3.0 – 2.9)

[1] First number represents actual number, number in parentheses is percentage of sample
[2] First number is mean score, number in parentheses is standard deviation. Range below these numbers is score range.

Table 58

Descriptive Characteristics for FAA Field Facility Sample (Original *N*=500)

Variable	Validation Sample	
	Developmental (*n*=318)	Cross-validation (*n*=182)
ATCS Cert./Wash Out[1]	**286** (89.9) / **32** (10.1)	**170** (93.4) / **12** (6.6)
Academy Final Score[2]	**79.1 (5.5)** (70.1 – 92.8)	**78.8 (5.4)** (70.2 – 94.0)
ATCS Entrance Exam[2]	**90.7 (5.1)** (73.2 – 100.0)	**91.0 (4.8)** (79.5 – 100.0)
Biodata Score (empirical)[2]	**96.82 (18.6)** (51.0 – 151.0)	**98.92 (17.7)** (58.0 – 144.0)
Factor Analytic Scores[2]		
General Expectations	**.02 (1.0)** (-2.9 - 2.4)	**-.03 (.9)** (-2.7 – 2.0)
Work Ethic/Personal Characteristics	**-.00 (1.0)** (-3.6 - 2.2)	**.01 (.9)** (-2.1 – 2.0)
Prior air traffic/Military Experience	**-.00 (1.0)** (-1.6 – 2.5)	**.01 (.9)** (-1.5 – 2.2)
High school academics	**-.02 (1.0)** (-2.9 - 2.3)	**.03 (1.0)** (-2.2 – 1.8)
High school athletics	**-.00 (1.0)** (-2.3 – 2.9)	**.00 (1.0)** (-2.9 – 2.3)

[1] First number represents actual number, number in parentheses is percentage of sample
[2] First number is mean score, number in parentheses is standard deviation. Range below these numbers is score range.

Table 59

Descriptive Characteristics for FAA Field Facility Sample (Original *N*=100)

Variable	Validation Sample	
	Developmental (*n*=67)	Cross-validation (*n*=33)
ATCS Cert./Wash Out[1]	**60** (89.6) / **7** (10.4)	**30** (90.9) / **3** (9.1)
Academy Final Score[2]	**78.8 (5.6)**	**78.6 (6.1)**
	(70.3 – 92.1)	(70.2 – 90.5)
ATCS Entrance Exam[2]	**91.6 (4.6)**	**90.1 (5.3)**
	(80.4 – 100.0)	(77.9 – 100.0)
Biodata Score (empirical)[2]	**189.85 (36.3)**	**185.55 (27.3)**
	(100.0 – 256.0)	(146.0 – 253.0)
Factor Analytic Scores[2]		
General Expectations	**.08 (1.0)**	**-.15 (.8)**
	(-2.8 - 2.7)	(-1.6 – 1.7)
Work Ethic/Personal	**.11 (1.0)**	**-.22 (.8)**
Characteristics	(-2.0 - 2.1)	(-2.3 – 1.0)
Prior air traffic/Military	**-.06 (1.0)**	**.13 (.8)**
Experience	(-1.5 – 1.9)	(-1.7 – 1.7)
Interpersonal affil.	**.09 (1.0)**	**-.18 (.6)**
	(-2.2 - 2.5)	(-1.3 – 1.3)

[1] First number represents actual number, number in parentheses is percentage of sample
[2] First number is mean score, number in parentheses is standard deviation. Range below these numbers is score range.

Table 60

Intercorrelations of Relevant Variables for FAA Field Facility Sample (Original *N*=2,000)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. OJT Status														
2. Academy	.03													
3. ATCS Exam	-.04	.15**												
4. Biodata	.18**	-.05	-.08**											
5. FA-GE	.06*	-.08**	-.07*	.23**										
6. FA-WE	.07*	.05	-.01	.11**	.01									
7. FA-PE	.02	.07*	-.18**	.08**	-.00	.00								
8. FA-HSAC	-.08**	.11**	.14**	-.05	.00	.00	.01							
9. FA-HSAT	.07*	-.04	-.04	.31**	-.01	.02	.00	-.00						
10. FA-ATT	.09**	.03	-.00	.19**	.00	-.00	.01	-.00	-.02					
11. MI-HSAC	-.06*	.12**	.16**	-.04	-.08**	.16**	-.20**	.92**	.02	.03				
12. MI-WE	.06*	.05	-.02	.13**	.11**	.93**	.00	.10**	.01	.06*	.24**			
13. MI-JS	.07*	-.07*	-.02*	.11**	.48**	.12**	-.05	-.08**	-.03	-.11**	-.07*	.14**		
14. MI-GE	.10**	-.08**	-.10**	.30**	.83**	.06*	.07*	-.10**	.06*	.28**	-.13**	.18**	.42**	
15. MI-PE	.04	.03	-.21**	.12**	.11**	-.01	.89**	-.24**	.02	.08**	-.40**	.00	.01	.23**

Sample n = 1,329; * $p \leq .05$ *(2-tailed)*, ** $p \leq .01$ *(2-tailed)*
Values in lower triangle are uncorrected for range restriction.
Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice, ATT=Attendance.

Table 61

Intercorrelations of Relevant Variables for FAA Field Facility Sample (Original *N*=1,500)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. OJT Status														
2. Academy	.05													
3. ATCS Exam	-.06	.16**												
4. Biodata	.18**	.03	-.12**											
5. FA-GE	.10**	-.05	-.08*	.11**										
6. FA-WE	.06	.06	.01	.11**	-.01									
7. FA-HSAC	-.05	.10**	.13**	-.21**	-.00	-.01								
8. FA-PE	.02	.06	-.18**	.28**	.01	.01	.01							
9. FA-HSAT	.12**	-.02	-.03	-.05	-.02	.02	.01	-.01						
10. FA-JS	-.02	.01	.07*	.02	.01	.00	.02	-.01	.02					
11. MI-HSAC	-.04	.10**	.16**	-.23**	-.06	.11**	.95**	-.16**	.05	.01				
12. MI-WE	.03	.07*	.02	.06	.03	.89**	.11**	.00	.05	.11**	.22**			
13. MI-GE	.13**	-.07*	-.07*	.13**	.84**	.11**	-.08*	-.03	.09**	.02	-.08*	.12**		
14. MI-JS	.08*	-.01	.02	.04	.35**	.06	-.07*	-.07*	.05	.65**	-.07*	.11**	.40**	
15. MI-PE	.06	-.01	-.21**	.38**	.04	.00	-.32**	.83**	-.01	.00	-.46**	-.05	.07*	-.02

Sample n = 997; * *p≤.05 (2-tailed),* ** *p≤.01 (2-tailed)*

Values in lower triangle are uncorrected for range restriction.

Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Table 62

Intercorrelations of Relevant Variables for FAA Field Facility Sample (Original *N*=1,000)

	1	2	3	4	5	6	7	8	9
1. OJT Status									
2. Academy	.02								
3. ATCS Exam	-.01	.19**							
4. Biodata	.34**	.01	-.10*						
5. FA-GE	.11**	-.06	.01	.35**					
6. FA-WE	.08	.05	-.01	.33**	-.02				
7. FA-PE	.04	.09*	-.20**	.26**	-.01	.02			
8. FA-HSAC	-.02	.15**	.12**	-.03	.02	.02	.00		
9. FA-HSAT	.07	-.07	-.09*	.07	-.01	.02	.00	-.00	
10. FA-ATT	.02	.02	-.04	.00	-.02	-.02	.00	.00	-.02

Sample n = 688; * $p \leq .05$ *(2-tailed)*, ** $p \leq .01$ *(2-tailed)*
Values in lower triangle are uncorrected for range restriction.
Biodata = Empirical Key, FA = Factor Analytic Key.
FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics,
PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics,
ATT=Attendance.

Table 63

Intercorrelations of Relevant Variables for FAA Field Facility Sample (Original *N*=500)

	1	2	3	4	5	6	7	8
1. OJT Status								
2. Academy	.07							
3. ATCS Exam	-.10	.09						
4. Biodata	.31**	.00	-.21**					
5. FA-GE	.08	-.00	.07	.11				
6. FA-WE	.05	.12*	.01	.16**	-.02			
7. FA-PE	.06	.16**	-.24**	.39**	.00	.04		
8. FA-HSAC	-.16**	.05	.19**	-.31**	.01	.01	.04	
9. FA-HSAT	.13*	-.01	.00	.00	.01	.04	-.00	-.02

Sample n = 318; * $p \leq .05$ *(2-tailed)*, ** $p \leq .01$ *(2-tailed)*
Values in lower triangle are uncorrected for range restriction.
Biodata = Empirical Key, FA = Factor Analytic Key.
FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics,
PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics.

Table 64

Intercorrelations of Relevant Variables for FAA Field Facility Sample (Original *N*=100)

	1	2	3	4	5	6	7
1. OJT Status							
2. Academy	-.14						
3. ATCS Exam	-.24	.11					
4. Biodata	.79**	-.16	-.42**				
5. FA-WE	.13	.03	-.14	.31*			
6. FA-GE	.23	-.21	-.27*	.34**	.03		
7. FA-PE	.18	.12	-.25*	.41**	-.01	.02	
8. FA-INT	.26*	.05	-.10	.25*	-.01	-.05	-.01

Sample $n = 67$; * $p \leq .05$ *(2-tailed)*, ** $p \leq .01$ *(2-tailed)*
Values in lower triangle are uncorrected for range restriction.
Biodata = Empirical Key, FA = Factor Analytic Key. FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, INT=Interpersonal Affiliation.

Table 65

Hierarchical Logistic Regression Analysis for prediction of FAA Field Facility pass/fail status from ATCS Entrance Exam, FAA Academy Score and Biodata Keys – Original N = 2,000

Step and Variable	e^B	Wald χ^2	-2LL	Δ-2LL	[1]Model R^2
Step 1					
ATCS Entrance Exam	.97	2.85	760.57	[2]3.98	.007
FAA Academy Score	1.02	1.63			
Step 2a (Empirical)					
ATCS Entrance Exam	.97	1.55	725.34	35.23***	.068
FAA Academy Score	1.03	2.08			
Biodata Key	1.58	34.96***			
Step 2b (FA)					
ATCS Entrance Exam	.98	.98	725.67	34.90***	.067
FAA Academy Score	1.03	2.30			
FA-GE	1.24	4.90*			
FA-WE	1.27	5.99*			
FA-PE	1.04	.10			
FA-HSAC	.74	8.49**			
FA-HSAT	1.28	6.32*			
FA-ATT	1.35	10.14**			
Step 2c (MIRT)					
ATCS Entrance Exam	.98	1.44	740.24	20.33**	.042
FAA Academy Score	1.03	2.67			
MI-HSAC	.73	4.19*			
MI-WE	1.29	3.18			
MI-JS	1.18	1.12			
MI-GE	1.48	5.24*			
MI-PE	.92	.21			

Sample n = 1,329; *$p \leq .05$ (2-tailed), **$p \leq .01$ (2-tailed), ***$p \leq .001$*

Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice, ATT=Attendance.

Notes: (1) Model R^2 as defined in Nagelkerke (1991); (2) Δ-2LL for Step 1 represents the difference between the –2LL for the model with no independent variables and that for the base model with two (ATCS Entrance Exam and FAA Academy Score) independent variables.

Table 66

Hierarchical Logistic Regression Analysis for prediction of FAA Field Facility pass/fail status from ATCS Entrance Exam, FAA Academy Score and Biodata Keys – Original N = 1,500

Step and Variable	e^B	Wald χ^2	-2LL	Δ-2LL	[1]Model R^2
Step 1					
ATCS Entrance Exam	.95	4.10*	570.17	[2]6.30*	.015
FAA Academy Score	1.04	3.05			
Step 2a (Empirical)					
ATCS Entrance Exam	.97	1.41	528.73	47.74***	.109
FAA Academy Score	1.03	2.38			
Biodata Key	1.12	25.08***			
Step 2b (FA)					
ATCS Entrance Exam	.97	1.84	541.41	35.06***	.080
FAA Academy Score	1.04	3.78			
FA-GE	1.38	8.77**			
FA-WE	1.23	3.17			
FA-HSAC	.84	2.23			
FA-PE	1.05	.13			
FA-HSAT	1.55	13.21***			
FA-JS	.92	.56			
Step 2c (MIRT)					
ATCS Entrance Exam	.96	2.61	552.14	24.33**	.056
FAA Academy Score	1.04	3.79			
MI-HSAC	.93	.15			
MI-WE	1.09	.29			
MI-GE	1.80	9.54**			
MI-JS	1.22	.76			
MI-PE	1.30	.79			

Sample n = 997; *$p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$*

Biodata = Empirical Key, FA = Factor Analytic Key, MI = Multidimensional IRT Key. FA and MI subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, JS=Job Security as motivator of job choice.

Notes: (1) Model R^2 as defined in Nagelkerke (1991); (2) Δ-2LL for Step 1 represents the difference between the -2LL for the model with no independent variables and that for the base model with two (ATCS Entrance Exam and FAA Academy Score) independent variables.

Table 67

Hierarchical Logistic Regression Analysis for prediction of FAA Field Facility pass/fail status from ATCS Entrance Exam, FAA Academy Score and Biodata Keys – Original N = 1,000

Step and Variable	e^B	Wald χ^2	-2LL	Δ-2LL	[1]Model R^2
Step 1					
ATCS Entrance Exam	.99	.20	378.11	[2].40	.001
FAA Academy Score	1.01	.28			
Step 2a (Empirical)					
ATCS Entrance Exam	1.02	.25	309.03	69.48***	.231
FAA Academy Score	1.01	.06			
Biodata Key	1.42	56.31***			
Step 2b (FA)					
ATCS Entrance Exam	1.00	.01	361.95	16.56*	.057
FAA Academy Score	1.02	.41			
FA-GE	1.49	7.91**			
FA-WE	1.31	3.67			
FA-PE	1.16	.88			
FA-HSAC	.88	.76			
FA-HSAT	1.30	3.08			
FA-ATT	1.08	.31			

Sample n = 688; *$p \leq .05$ (2-tailed), **$p \leq .01$ (2-tailed), ***$p \leq .001$

Biodata = Empirical Key, FA = Factor Analytic Key.

FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics, ATT=Attendance.

Notes: (1) Model R^2 as defined in Nagelkerke (1991); (2) Δ-2LL for Step 1 represents the difference between the -2LL for the model with no independent variables and that for the base model with two (ATCS Entrance Exam and FAA Academy Score) independent variables.

Table 68

Hierarchical Logistic Regression Analysis for prediction of FAA Field Facility pass/fail status from ATCS Entrance Exam, FAA Academy Score and Biodata Keys – Original $N = 500$

Step and Variable	e^B	Wald χ^2	-2LL	Δ-2LL	[1]Model R^2
Step 1					
ATCS Entrance Exam	.92	3.36	170.04	[2]5.16	.039
FAA Academy Score	1.06	1.98			
Step 2a (Empirical)					
ATCS Entrance Exam	.94	1.11	128.56	46.64***	.327
FAA Academy Score	1.04	.79			
Biodata Key	1.14	25.15***			
Step 2b (FA)					
ATCS Entrance Exam	.95	1.31	155.47	19.73**	.145
FAA Academy Score	1.05	1.59			
FA-GE	1.35	2.08			
FA-WE	1.16	.46			
FA-PE	1.30	1.01			
FA-HSAC	.55	6.23*			
FA-HSAT	1.58	4.14*			

Sample $n = 318$; $p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$
Biodata = Empirical Key, FA = Factor Analytic Key.
FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, HSAC=High School Academics, HSAT=High School Athletics.
Notes: (1) Model R^2 as defined in Nagelkerke (1991); (2) Δ-2LL for Step 1 represents the difference between the –2LL for the model with no independent variables and that for the base model with two (ATCS Entrance Exam and FAA Academy Score) independent variables.

Table 69

Hierarchical Logistic Regression Analysis for prediction of FAA Field Facility pass/fail status from ATCS Entrance Exam, FAA Academy Score and Biodata Keys – Original $N = 100$

Step and Variable	e^B	Wald χ^2	-2LL	Δ-2LL	[1]Model R^2
Step 1					
ATCS Entrance Exam	.86	2.85	39.68	[2]4.30	.131
FAA Academy Score	.93	.85			
Step 2a (Empirical)[3]					
ATCS Entrance Exam					
FAA Academy Score					
Biodata Key					
Step 2b (FA)					
ATCS Entrance Exam	.94	.25	29.62	10.06*	.405
FAA Academy Score	.88	1.80			
FA-WE	2.12	1.67			
FA-GE	2.77	2.07			
FA-PE	1.77	.83			
FA-INT	3.28	3.80			

Sample $n = 67$; $p \leq .05$ (2-tailed), ** $p \leq .01$ (2-tailed), ***$p \leq .001$
Biodata = Empirical Key, FA = Factor Analytic Key.
FA subkeys: GE=General Expectations, WE=Work Ethic/Personal Characteristics, PE=Previous Air Traffic/Military Experience, INT=Interpersonal Affiliation.
Notes: (1) Model R^2 as defined in Nagelkerke (1991); (2) Δ-2LL for Step 1 represents the difference between the –2LL for the model with no independent variables and that for the base model with two (ATCS Entrance Exam and FAA Academy Score) independent variables; (3) Model convergence was not reached for this analysis.

Table 70

FAA Field Facility (Pass/Fail) Validities and Cross-Validities as a Function of Keying Method and Sample Size

	Samples									
	2000		1500		1000		500		100	
	V	CV	V	CV	V	CV	V	CV	V	CV
Sample Size	1329	671	997	503	688	312	318	182	67	33
ATCS Exam and Academy Score	.06	.13	.09	.04	.03	.14	.10	-.04	**.31**[1]	**-.26**[1]
Exam/Academy + Biodata										
Empirical	**.17**[2]	**.06**[2]	**.24**[1]	**.02**[1]	**.35**[1]	**.05**[1]	**.49**[1]	**-.08**[1]	*[4]	*[4]
FA	.17	.17	**.23**[2]	**.11**[2]	.17	.16	**.27**[1]	**-.12**[1]	**.52**[1]	**-.23**[1]
MIRT	.15	.17	.18	.08	*[3]	*[3]	*[3]	*[3]	*[3]	*[3]

FAA Academy pass/fail status used as criterion. Logistic regression weights applied.
Note: V = validity; CV = cross-validity. (1) Difference between validity and cross-validity is significant at $p \leq .01$; (2) Difference between validity and cross-validity is significant at $p \leq .05$; (3) No keys using the MIRT were available at this sample size. (4) Convergence never reached in logistic regression with empirical key added as a predictor.

Table 71

FAA Field Facility (Pass/Fail Status) Cross-Validities as a Function of Keying Method and Sample Size

	Samples				
	2000	1500	1000	500	100
Sample Size	671	503	312	182	33
ATCS Exam and Academy Score	.13**	.04	.14*	-.04	-.26
Exam/Academy + Biodata					
Empirical	.06	.02	.05	-.08	*[2]
FA	.17**	.11*	.16**	-.12	-.23
MIRT	.17**	.08	*[1]	*[1]	*[1]

* $p \leq .05$; ** $p \leq .01$

FAA Academy pass/fail status used as criterion. Logistic regression weights applied.

Note: (1) No keys using the MIRT were available at this sample size.
(2) Convergence never reached in logistic regression with emp.rical key added as a predictor.

Table 72

FAA Field Facility (Pass/Fail Status) Classification Accuracy as a Function of Keying Method and Sample Size

	Samples				
	2000	1500	1000	500	100
Sample Size	671	503	312	182	33
ATCS Exam and Academy Score	91.1	89.9	91.5	94.0	90.0
Exam/Academy + Biodata					
Empirical	91.1	89.9	89.8	93.4	*[2]
FA	91.1	89.9	91.5	94.0	87.9
MIRT	91.1	89.9	*[1]	*[1]	*[1]

FAA Academy pass/fail status used as criterion. Logistic regression weights applied to individual exam, Academy final scores and biodata scores to calculate predicted probability of passing. Number represents percentage of individuals correctly classified.
Note: (1) No keys using the MIRT were available at this sample size.
(2) Convergence never reached in logistic regression with empirical key added as a predictor.

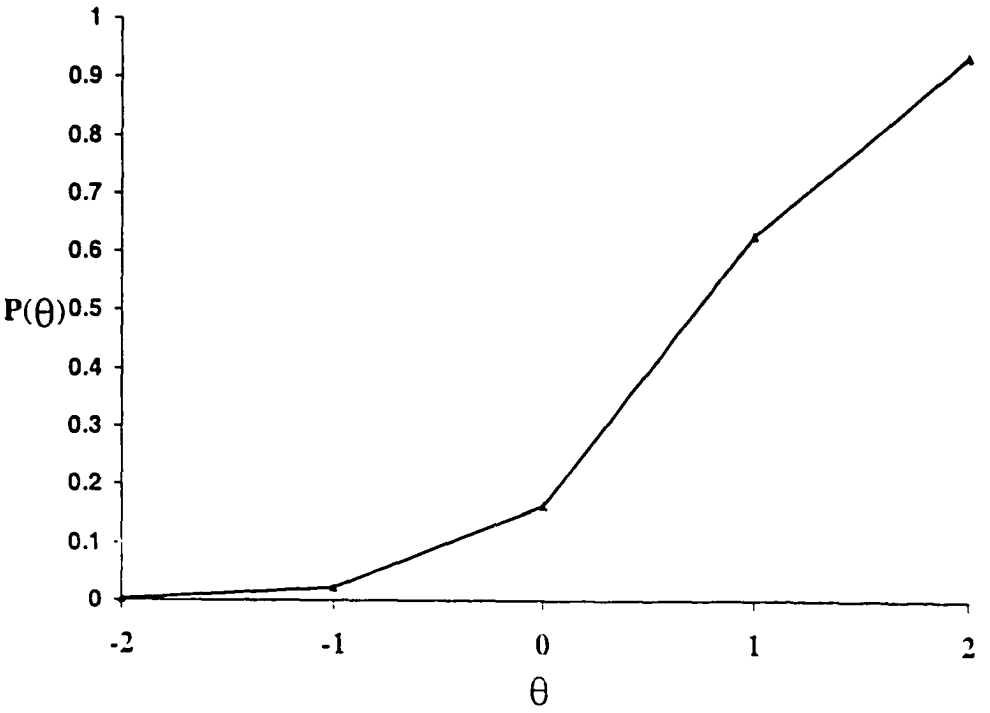
Figure Captions

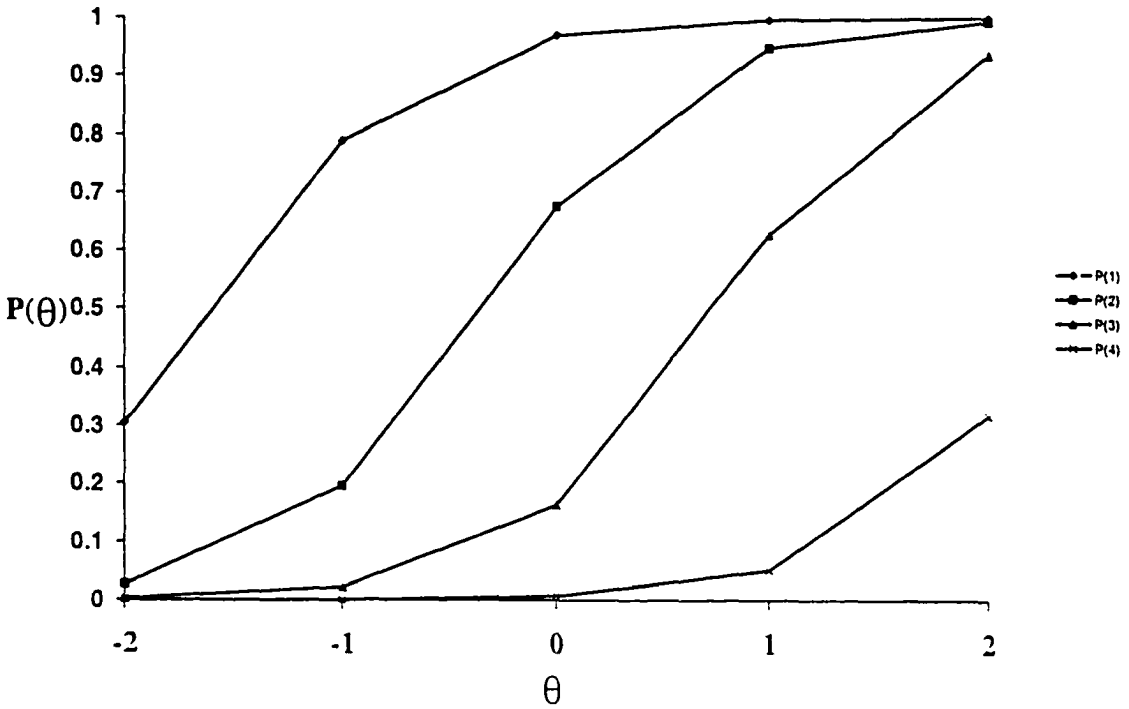
Figure 1. Item characteristic curve (ICC) for binary-response item.

Figure 2. Item characteristic curves (ICCs) for graded-response (5 possible response options) item. Each curve represents the probability of selecting a response category over the one immediately preceding it, along the response continuum.

Figure 3. Sample Multiplex Controller Aptitude Test (MCAT) item.

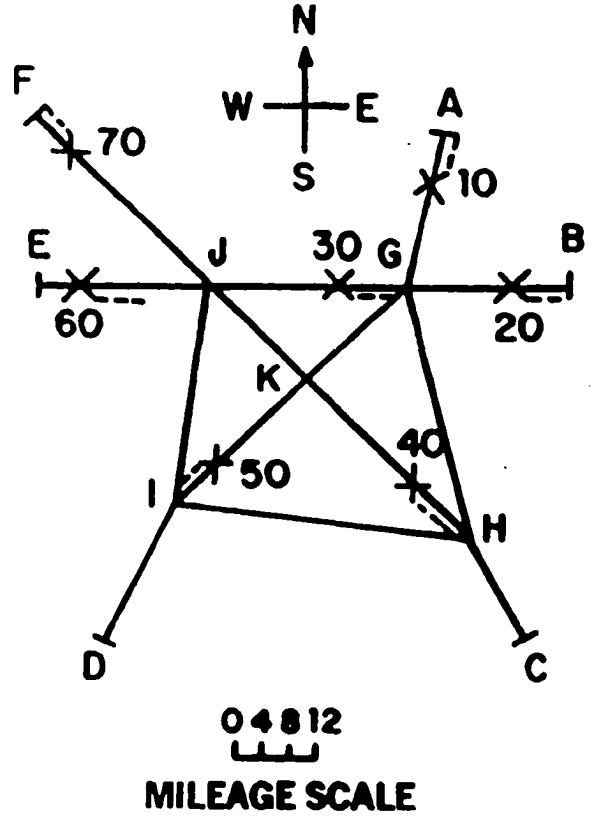
Figure 4. Sample Abstract Reasoning Test (ABSR) item.


1
0.9
0.8
0.7
0.6
P(θ) 0.5
0.4
0.3
0.2
0.1
0
-2
-1
0
1
2
θ


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P(θ)
-2
-1
0
1
2
θ
P(1)
P(2)
P(3)
P(4)

MULTIPLEX CONTROLLER APTITUDE TEST (MCAT)

AIRCRAFT	ALTITUDE	SPEED	ROUTE
10	7000	480	AGKHC
20	7000	480	BGJE
30	7000	240	AGJE
40	6500	240	CHKJF
50	6500	240	DIKGB
60	8000	480	DIKJE
70	8000	480	FJKID



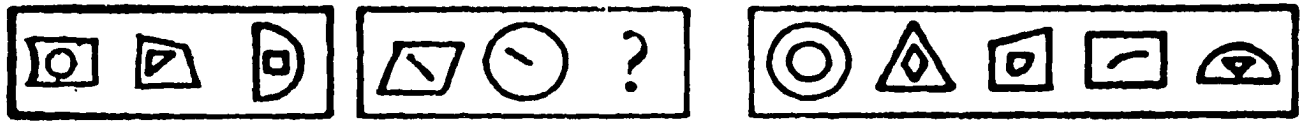
SAMPLE QUESTION

WHICH AIRCRAFT WILL CONFLICT?

A. 60 AND 70
B. 40 AND 70
C. 20 AND 30
D. NONE OF THESE

ABSTRACT REASONING

Symbols

1. 

2. 

Letters

1) XCXDXEX A) FX B) FG C) XF D) EF E) XG

2) ARCSETG A) HI B) HU C) UJ D) UI E) IV

APPENDIX A
Biographical Questionnaire (BQ)

INSTRUCTIONS: ALL the items, which follow, are in a multiple-choice format. Answer each one by blackening the oval in the appropriate column of your choice. Choose the response that best fits you and **MAKE ONLY ONE RESPONSE PER QUESTION.**

EDUCATION

* 1. Which of the following best describes your PRIMARY/SECONDARY education?
No formal education
Elementary school
Some high school
High school graduate
High school GED

* 2. Which of the following best describes your TECHNICAL SCHOOL education?
None/Not applicable
Military - not completed
Military - completed
Vocational/Technical - not completed
Vocational/Technical - completed

3. Which of the following best describes the amount of UNDERGRADUATE COLLEGE education that you have completed?
None/Not applicable
Less than one year
At least one year, but less than 2 years
At least two years, but less than 3 years
Three or more years

4. Which of the following best describes POSTGRADUATE COLLEGE course work that you have taken?
None/Not applicable
Master's degree course work
Professional degree course work (law, medicine)
Doctoral degree course work
Post-Doctoral course work

5. Please indicate the HIGHEST college degree that you have received.
None/Not applicable
Associate degree
Bachelor's degree
Master's degree
Professional or Doctoral degree

Social Security Number

Example: 999 99 9999

What grades, on the average, did you get in the following HIGH SCHOOL courses? Fill in the oval corresponding to the average grade for each subject.

6. Arithmetic/Math (give average of all math courses combined)
About "A-" to "A+"
About "B-" to "B+"
About "C-" to "C+"
Lower than "C-"
Did not have the course

7. Physical Sciences (give average of all physical sciences courses combined)
About "A-" to "A+"
About "B-" to "B+"
About "C-" to "C+"
Lower than "C-"
Did not have the course

8. English
About "A-" to "A+"
About "B-" to "B+"
About "C-" to "C+"
Lower than "C-"
Did not have the course

* 9. Your overall high school average was:
About "A-" to "A+"
About "B-" to "B+"
About "C-" to "C+"
Lower than "C-"
Did not have the course

* 10. If you attended college, please indicate your OVERALL grade point average, computed on a 4-point scale.
If you did not attend college, please skip to item 11.
2.00 or below
2.01 to 2.50
2.51 to 3.00
3.01 to 3.50
3.51 to 4.00

* 11. How long has it been since you last attended school as a full-time student?
Less than one year
At least one year, but less than 2 years
At least two years, but less than 3 years
At least three years, but less than 4 years
Four years or more

Please indicate the number of college credits in semester hours that you earned in the following subject areas.

* 12. Accounting/Bookkeeping/Business/Finance/Marketing
Did not attend college
Attended college, earned 0 credit hours in this area of study
1 - 6 credit hours
7 - 12 credit hours
13 or more credit hours

* 13. Agriculture/Home Economics
Did not attend college
Attended college, earned 0 credit hours in this area of study
1 - 6 credit hours
7 - 12 credit hours
13 or more credit hours

* 14. Art/Music/Dance/Drama
Did not attend college
Attended college, earned 0 credit hours in this area of study
1 - 6 credit hours
7 - 12 credit hours
13 or more credit hours

* 15. Botany/Biology
Did not attend college
Attended college, earned 0 credit hours in this area of study
1 - 6 credit hours
7 - 12 credit hours
13 or more credit hours

* 16. Computer Science
Did not attend college
Attended college, earned 0 credit hours in this area of study
1 - 6 credit hours
7 - 12 credit hours
13 or more credit hours

* 17. Education
Did not attend college
Attended college, earned 0 credit hours in this area of study
1 - 6 credit hours
7 - 12 credit hours
13 or more credit hours

* 18. Engineering/Architecture
Did not attend college
Attended college, earned 0 credit hours in this area of study
1 - 6 credit hours
7 - 12 credit hours
13 or more credit hours

* 19. Foreign Languages
Did not attend college Attended college, earned 0 credit hours in this area of study 1 - 6 credit hours 7 - 12 credit hours 13 or more credit hours

* 20. Geology/Chemistry/Physics/Physical Science
Did not attend college Attended college, earned 0 credit hours in this area of study 1 - 6 credit hours 7 - 12 credit hours 13 or more credit hours

* 21. Humanities/English
Did not attend college Attended college, earned 0 credit hours in this area of study 1 - 6 credit hours 7 - 12 credit hours 13 or more credit hours

* 22. Management
Did not attend college Attended college, earned 0 credit hours in this area of study 1 - 6 credit hours 7 - 12 credit hours 13 or more credit hours

* 23. Mathematics/Statistics
Did not attend college Attended college, earned 0 credit hours in this area of study 1 - 6 credit hours 7 - 12 credit hours 13 or more credit hours

* 24. Psychology/Sociology/History/Human Relations
Did not attend college Attended college, earned 0 credit hours in this area of study 1 - 6 credit hours 7 - 12 credit hours 13 or more credit hours

* 25. Pre-professional (Pre-Med, Pre-Law, etc.)
Did not attend college Attended college, earned 0 credit hours in this area of study 1 - 6 credit hours 7 - 12 credit hours 13 or more credit hours

* 26. Speech/Journalism/Communications
Did not attend college Attended college, earned 0 credit hours in this area of study 1 - 6 credit hours 7 - 12 credit hours 13 or more credit hours
27. Which of the following general categories BEST describes aviation course work taken toward your Associate and/or Technical/Military/Vocational-Technical degree(s)? Please indicate only one response.
No aviation course work/Not applicable Vocational oriented (Training directed toward a specific occupation and/or FAA certification(s) rather than toward further education in a Bachelor's degree program) Baccalaureate-transfer oriented (Training directed toward further education in a Bachelor's degree program rather than toward a specific occupation or FAA certificate) Other

28. Which of the following categories BEST describes aviation course work taken toward your Bachelor's degree? Please indicate only one response.
No aviation course work/Not applicable Aviation Operations (Education focusing on the operation of aircraft on the ground or in the air; e.g., Flight Engineer, Air Traffic Control, Professional Pilot) Aviation Technology (Education focusing on ground support function; e.g., Avionics, Aviation Maintenance) Aviation Management (Education focusing on the management of personnel and/or operations or systems; e.g., Aviation Administration/Management, Air Transportation Management, Airline Management) Aviation-Other (e.g., Airway Computer Science, Aviation/Aerospace/Aeronautical Engineering)
* 29. How many aviation-related college credits did you earn? (Credit hours received from both 4-year and 2-year institutions can be included, but those earned at technical schools in programs not leading to a degree should not be included.)
0 - 6 credit hours (Includes those who did not attend college) 7 - 14 credit hours 15 - 22 credit hours 23 - 29 credit hours 30 or more credit hours

MILITARY EXPERIENCE

* 30. Have you ever been in the Armed Services?
No - (Skip item 31) Yes - Military Reserve Yes - Regular Service
* 31. Which branch of the service?
United States Army United States Air Force United States Coast Guard United States Marine Corps United States Navy

* 32. Did you work in an aviation-related or air traffic-related job while in the armed forces?
Was not in armed forces Did not work in aviation-related job while in armed forces Worked in aviation-related job which was not related to air traffic Worked in air traffic-related job
33. Do you have a prior Control Tower Operator (CTO) rating?
No Yes - IFR Yes - VFR Yes - Both IFR and VFR
34. Do you have a prior Air Traffic Control Specialist (ATCS) rating?
No Yes - Center Yes - Flight Service Station (FSS/AFSS) Yes - TRACON/RAPCON Yes - Center, FSS/AFSS, and TRACON/RAPCON
35. Do you have prior IFR operations experience?
No Yes - Military Yes - Civilian Yes - Both Military and Civilian

For items 36 through 42, estimate the amount of IFR experience you have had.

36. ARTCC
None
Under 6 months
6 to under 12 months
12 to under 18 months
18 months or over

37. RATCC OR CATCC
None
Under 6 months
6 to under 12 months
12 to under 18 months
18 months or over

38. ARAC
None
Under 6 months
6 to under 12 months
12 to under 18 months
18 months or over

39. RAPCON
None
Under 6 months
6 to under 12 months
12 to under 18 months
18 months or over

40. TOWER
None
Under 6 months
6 to under 12 months
12 to under 18 months
18 months or over

41. GCA (RADAR)
None
Under 6 months
6 to under 12 months
12 to under 18 months
18 months or over

42. GCI (RADAR)
None
Under 6 months
6 to under 12 months
12 to under 18 months
18 months or over

43. Have you had prior VFR operations experience?
No
Yes - Military
Yes - Civilian
Yes - Both Military and Civilian

For items 44 through 46, estimate the amount of VFR experience you have had.

44. Tower
None
Under 6 months
6 to under 12 months
12 to under 18 month
18 months or over

45. FSS or IFSS
None
Under 6 months
6 to under 12 months
12 to under 18 month
18 months or over

46. GCI (NON-Radar)
None
Under 6 months
6 to under 12 months
12 to under 18 month
18 months or over

COMMUNICATIONS EXPERIENCE
For items 47 through 54, estimate the amount of communications OPERATIONS (not maintenance) experience you have had.

47. Citizens Band (CB)
None
Less than 1 year
At least one year but less than 2 years
At least two years but less than 3 years
3 years or more

48. Ham radio operator
None
Less than 1 year
At least one year but less than 2 years
At least two years but less than 3 years
3 years or more

49. Air-to-air
None
Less than 1 year
At least one year but less than 2 years
At least two years but less than 3 years
3 years or more

50. Air-to-ground
None
Less than 1 year
At least one year but less than 2 years
At least two years but less than 3 years
3 years or more

51. Point-to-point
None
Less than 1 year
At least one year but less than 2 years
At least two years but less than 3 years
3 years or more

52. Ship-to-shore
None
Less than 1 year
At least one year but less than 2 years
At least two years but less than 3 years
3 years or more

53. Ship-to-ship
None
Less than 1 year
At least one year but less than 2 years
At least two years but less than 3 years
3 years or more

54. Computer communications
None
Less than 1 year
At least one year but less than 2 years
At least two years but less than 3 years
3 years or more

RATING(S)/CERTIFICATE(S)/LICENSE(S) POSSESSED:
For items 55 through 71, indicate whether or not you currently, or have ever possessed any of the ratings/ certificates/licenses.

55. Student pilot
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

56. Private pilot
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

57. Commercial pilot
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

58. Airline transport pilot
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

59. Flight instructor
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

60. Instrument flight instructor
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

61. Ground instructor
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

62. Instrument ground instructor
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

63. Single-engine
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

64. Multi-engine
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

65. Instrument
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

66. Airplane
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

67. Rotorcraft
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

68. Glider
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

69. Lighter-than-air
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

70. Dispatch - air carrier
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

71. Navigator/bombardier
No rating, certificate or license
Military rating, certificate or license
Civilian rating, certificate or license
Both military and civilian rating, certificate or license

72. Indicate the PRIMARY method you used to prepare for the Office of Personnel Management (OPM) Air Traffic Control written test.
Did not prepare
Prepared by using the ATC test materials provided by OPM
Was coached by FAA employee(s) regarding the ATC selection tests
Prepared using published materials describing ATC selection tests
Attended workshop that taught ATC selection test-taking techniques, or used materials provided through a workshop

73. How many times did you take the OPM Air Traffic Controller Aptitude Test?
Never
Once
Twice
Three times
Four or more times

74. How did you enter Academy training?
Was selected competitively (off regular or Airway Science registers)
Entered through special hiring program (e.g., Pre-developmental, Cooperative Education, Upward Mobility, etc.)
Transferred from non-air traffic job with FAA or other government agency (e.g., FAA secretary, FAA electronics technician, former postal worker)
Transferred from other air traffic-related job not included in 2152 series (e.g., air traffic assistant, flight data specialist)
GS-2152 transferring to new option (En route, Terminal, FSS)

75. How were you interviewed for this job?
Spoke with a FAA representative by telephone
Spoke in person with a FAA representative at a FAA facility located at an airport or center
Spoke in person with a FAA representative at another type of FAA office
Spoke in person or by telephone with a non-FAA representative
Was interviewed by another method

For items 76 through 84, rate the EXTENT to which the following aspects of the ATC occupation were explained to you before you accepted this job.

76. Your assignment to a particular region
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

77. Your assignment to a particular option
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

78. Your assignment to a particular facility
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

79. Requirement to work shifts
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

80. Standards, which must be met to successfully, complete training
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

81. Your opportunities for career advancement
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

82. Your duties and benefits as an air traffic controller
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

83.Your duties and benefits as a government employee
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

84. Your alternatives as a government employee if you do not pass this program.
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

PRIOR FAA EXPERIENCE

85. How many times have you previously attended an En route or Terminal non-radar-screening course at the Academy?
Never
Once before
Twice before
Three time before
Four or more times before

86. What was your status at the end of the last Academy non-radar screening course you attended?
Did not previously attend
Attended, Passed
Attended, Failed
Attended, Withdrew

87. If you previously attended an Academy non-radar course but did not pass, what condition allowed you to reenter?
Not applicable
Was rehired competitively (through register)
Obtained additional aviation-related experience to reenter
Was allowed to reenter due to extenuating circumstances
Reentered as result of legal action

OCCUPATIONAL CHOICE

How IMPORTANT was each of the following factors in influencing your choice of the ATC occupation?

* 88. Salary
Not at all important
Of limited importance
Of moderate importance
Of considerable importance
Of very great importance

* 89. Benefits
Not at all important
Of limited importance
Of moderate importance
Of considerable importance
Of very great importance

* 90. Job security
Not at all important
Of limited importance
Of moderate importance
Of considerable importance
Of very great importance

* 91. Importance of the job
Not at all important
Of limited importance
Of moderate importance
Of considerable importance
Of very great importance

* 92. Opportunity to work with hands
Not at all important
Of limited importance
Of moderate importance
Of considerable importance
Of very great importance

* 93. Intellectual challenge
Not at all important
Of limited importance
Of moderate importance
Of considerable importance
Of very great importance

* 94. Interest in aviation
Not at all important
Of limited importance
Of moderate importance
Of considerable importance
Of very great importance

* 95. Opportunity for advancement
Not at all important
Of limited importance
Of moderate importance
Of considerable importance
Of very great importance

* 96. Autonomy (ability to work independently)
Not at all important
Of limited importance
Of moderate importance
Of considerable importance
Of very great importance

* 97. Prestige of the job
Not at all important
Of limited importance
Of moderate importance
Of considerable importance
Of very great importance

* 98. Opportunity to work with competent people
Not at all important
Of limited importance
Of moderate importance
Of considerable importance
Of very great importance

* 99. Prior aviation-related experience
Not at all important
Of limited importance
Of moderate importance
Of considerable importance
Of very great importance

* 100. Ability to control workload
Not at all important
Of limited importance
Of moderate importance
Of considerable importance
Of very great importance

* 101. Opportunity to be admired by others
Not at all important
Of limited importance
Of moderate importance
Of considerable importance
Of very great importance

* 102. Opportunity to benefit others
Not at all important
Of limited importance
Of moderate importance
Of considerable importance
Of very great importance

* 103. High Salary
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 104. Good Benefits
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 105. Good Job security
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 106. Opportunity to work with hands
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 107. Intellectual challenge
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 108. Opportunity for advancement
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 109. Autonomy (ability to work independently)
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 110. Opportunity to work with competent people
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

PERFORMANCE EXPECTATIONS

* 111. How long do you think that it will take you to become fully effective in your current job?
Much longer than most others
Somewhat longer than most others
About as long as most others
A little less time than most others
Much less time than most others

* 112. Of all the air traffic controllers in the country, at what percentile do you think you will be able to perform?
In the lowest 10%
In the lower half
At about the 50% or average level
In the upper half
In the top 10%

GENERAL EXPECTATIONS

* 113. Do you believe that the FAA will continue to employ you if you perform well?
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 114. Is your career important to you?
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 115. Do you expect to be challenged by your job?
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 116. Do you expect your job to be equally challenging to you in five years?
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 117. Is it important for you to contribute to decisions affecting your job?
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 118. Do you understand what your future job duties will be?
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 119. Will it be difficult for you to adjust to a rotating work schedule?
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 120. Is it likely that increased automation will affect your job tasks and responsibilities?
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 121. Would you feel negatively if increased automation altered your job tasks and responsibilities?
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 122. Is this job important to you?
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 123. Is becoming a manager or supervisor important to you?
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 124. Do you believe that the FAA will promote those who are the best qualified?
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 125. Do you expect that management will be supportive of your concerns?
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 126. Do you expect that working for the Federal Government will be desirable?
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

* 127. Do you expect to be satisfied with your job?
Not at all
To a limited extent
To a moderate extent
To a considerable extent
To a very great extent

GENERAL BACKGROUND INFORMATION

* 128. Before your appointment as an Air Traffic Control Specialist, how often had you flown in an airplane?
Never
1 to 5 times
6 to 10 times
11 to 20 times

129. Have you ever visited a FAA center, tower cab or tower radar room?
No
Yes

* 130. While you were growing up, where did you live most of the time?
In a rural area
In a small town
In a suburb of a large city
In a city of less than 500,000 inhabitants
In a city of 500,000 or more inhabitants

* 131. While you were growing up, what was the economic status of your family?
Lower
Lower middle
Middle
Upper middle
Upper

* 132. Which of the following best describes your current marital status?
Single, never married
Separated
Divorced/widow/widower
Cohabiting ("living together")
Married

133. Which of the following best describes your current living arrangements?
Not married and not cohabiting
Spouse/"live-in" came to Oklahoma City and plans on, or is working outside the home
Spouse/"live-in" came to Oklahoma City and does not plan on working outside the home
Spouse/"live-in" stayed at prior residence and is working outside the home
Spouse/"live-in" stayed at prior residence and is not working outside the home

134. How much tobacco do you smoke?
Do not smoke cigarettes
Less than 1/2 pack per day
1/2 to 1 pack per day
More than 1, but less than 2 packs per day
2 packs or more per day

RELAXATION TECHNIQUES
Please indicate how often you engage in the following activities when you feel NERVOUS or TENSE.

135. Engage in a relaxation technique (yoga, meditation, etc.)
Almost never
Infrequently
Occasionally
Frequently
Almost always

136. Engage in physical activity (jogging, exercise program, etc.)
Almost never
Infrequently
Occasionally
Frequently
Almost always

137. Have an alcoholic drink
Almost never
Infrequently
Occasionally
Frequently
Almost always

138. Take a tranquilizer or other medication
Almost never
Infrequently
Occasionally
Frequently
Almost always

139. Engage in a passive activity (play or listen to music, read, etc.)
Almost never
Infrequently
Occasionally
Frequently
Almost always

140. Talk to a friend/co-worker/family member
Almost never
Infrequently
Occasionally
Frequently
Almost always

141. Use humor
Almost never
Infrequently
Occasionally
Frequently
Almost always

142. Eat/Eating
Almost never
Infrequently
Occasionally
Frequently
Almost always

143. How old were you when you had your first alcoholic beverage (including wine or beer) outside the home?
I never drink
Under age 14
Age 14 through 16
Age 17 through 20
Age 21 or older

144. How often during the past year did you have one or more alcoholic drinks?
I am a non-drinker
Once or twice a month
Once or twice a week
Most days
Every day

145. How many alcoholic drinks do you ordinarily consume at one sitting (from the time you start drinking until you quit)?
0 (non-drinker)
1 to 2
3 to 6
7 to 11
12 or more

THE FOLLOWING ITEMS WERE NOT INCLUDED IN THE SURVEY THAT WAS PRESENTED TO THE INDIVIDUALS THAT PROVIDED THE DATA FOR THIS STUDY

147.Do you consider yourself –?
Definitely, right handed
Somewhat right handed
Ambidextrous (use both hands with equal ease)
Somewhat left handed
Definitely left handed

TEST PREPARATION
We are attempting to understand, in more detail, how people go about preparing for the ATC written examination.

What kinds of materials or course(s), if any did you use to prepare for the OPM written air traffic control aptitude test?
Book or software
Self-study course (including book, practice problems, audio/video tapes, etc.)
Seminar or class offered by a company, organization, proprietary technical school, or person
College, university, junior college, or Vo-Tech school course or program
Did not use any materials or course(s) to prepare

How much did you pay for any materials or course(s) to prepare for the OPM written air traffic control aptitude test?
Nothing - materials or course(s) were offered free
Less than $50 total
Between $50 and $250, total
More than $250 total
Did not use any materials or course(s) to prepare

About how much time did you actually, spend in any course, class, or seminar that you attended to prepare for the ATC test?
About 4 hours or less
Between 4 and 8 hours
Between 8 and 16 hours (1 to 2 days)
More than 16 hours (3 or more days)
Did not attend a course, class, or seminar

Who sponsored any course, class, or seminar on the ATC test that you attended?
Organization such as the Black Coalition or Professional Women Controllers
FAA Region
College, university, junior college, or Vo-Tech school
Company, proprietary technical school, or person
Did not attend a course, class, or seminar

How did you find out about the course, class, or seminar on the ATC test?
Radio/TV advertisement
Magazine/newspaper advertisement
Mailed invitation
Personal contact
Did not attend a course, class, or seminar

What was the instructor's background in air traffic control in any course, class, or seminar on the ATC test you attended?
Current FAA employee
Former FAA employee
Other (Military, college teacher, etc.)
Didn't say or don't know
Did not attend a course, class, or seminar

What was the author's background in air traffic control for any book, software program, or self-study course that you used to prepare for the ATC test?
Current FAA employee
Former FAA employee
Other (Military, college teacher, etc.)
Didn't say or don't know
Did not use a book, software program, or self-study course

About how much time did you actually, spend using the materials from a book, software program, or self-study course to prepare for the ATC test?
About 8 hours or less
Between 8 and 16 hours (1 to 2 days) total
Between 16 and 40 hours (3 to 5 days) total
More than 40 hours (more than 5 days) total
Did not use a book, software program, or self-study course

RECRUITING
Your answers to the following questions will be used to help design ATC recruiting programs and to target advertising campaigns.

What was your PRIMARY OCCUPATION at the time you decided to apply for a FAA air traffic control specialist position?
Student
Employee for a private company or organization
Federal, state, or local government employee (not military)
Active-duty military service
Unemployed

About how much income did you earn during the past 12 months?
Less than $5,000
Between $5,000 and $15,000
Between $15,000 and $20,000
Between $20,000 and $25,000
More than $25,000

For items 157 through 168, indicate which SOURCE you got information from that made you decide to pursue a career as a FAA air traffic control specialist.

Newspaper/magazine article
No
Yes

Newspaper advertisement
No
Yes

Television/radio news story
No
Yes

Television commercial
No
Yes

College or university placement office
No
Yes

Government job listing
No
Yes

Radio commercial
No
Yes

Poster
No
Yes

FAA Recruiter
No
Yes

Relative
No
Yes

Friend
No
Yes

Other
No
Yes

Please indicate which type of MAGAZINES you read most frequently by blackening the oval in the appropriate column of your choice.

Aviation (e.g., Plane, Private Pilot, Aerospace Weekly)
No
Yes

Science (e.g., Omni, Scientific American, Discover)
No
Yes

Women's (e.g., Cosmopolitan, Ladies Home Journal)
No
Yes

Business (e.g., Business Week, Fortune, Forbes, Money)
No
Yes

Sports (e.g., Sports Illustrated, Golf Weekly)
No
Yes

Men's (e.g., Men's Health, GQ)
No
Yes

Computer (e.g., PC World, Byte, Compute, MacWorld)
No
Yes

News (e.g., Time, US News & World Report, Newsweek)
No
Yes

Minority (e.g., Jet, Ebony)
No
Yes

General interest (e.g., Life)
No
Yes

Please identify the types of TELEVISION PROGRAMS that you watch most frequently by blackening the oval in the appropriate column of your choice.

Comedy
No
Yes

Sports
No
Yes

Variety
No
Yes

Drama
No
Yes

Music video
No
Yes

News
No
Yes

Mystery
No
Yes

Nature/Science
No
Yes

Cartoons
No
Yes

Please indicate which types of RADIO PROGRAMMING you listen to most frequently by blackening the oval in the appropriate column of your choice.

Rock/Alternative
No
Yes

Easy listening
No
Yes

Classical
No
Yes

Religious
No
Yes

Soul
No
Yes

Golden oldies/classic rock
No
Yes

Sports
No
Yes

Talk
No
Yes

APPENDIX B
Applicant Background Assessment (ABA)

INSTRUCTIONS: Please answer all the questions on this biographical questionnaire to the best of you ability. **Your answers, which will be used for research purposes only and remain confidential,** will assist the Federal Aviation Administration Civil Aeromedical Institute (CAMI) in its longitudinal study of the Air Traffic selection process.

ALL the questions, which follow, are in a multiple-choice format. Answer each one by blackening the oval in the appropriate column of your choice. Choose the response that best fits you and **MAKE ONLY ONE RESPONSE PER QUESTION.**

ACADEMIC EXPERIENCE: HIGH SCHOOL	
1	During high school (grades 9-12) I made the semester honor roll: never once or twice three or four times five or six times seven or eight times
2	When I graduated from high school I was: 16 years old or younger 17 years old 18 years old 19 years old 20 years old or older
3	Relative to the other high school students in my major field of study, my most demanding teacher would most likely describe my academic work as: superior above average average below average don't know
4	During my last year in high school, my average number of hours of paid employment per week was: more than 20 16 to 20 hours 10 to 15 hours fewer than 10 hours none
5	Relative to the other high school students in my major field of study, my classmates would most likely describe my interpersonal skills as: superior above average average below average don't know
6	Relative to the other high school students in my major field of study, my classmates would most likely describe my leadership skills as: superior above average average below average don't know

Social Security Number

Example: **999** **99** **9999**

APPENDIX B
Applicant Background Assessment (ABA)

7	My high school teachers would most likely describe my self discipline as: superior above average average below average don't know
8	My high school teachers would most likely describe my academic potential as: superior above average average below average don't know
9	My high school classmates would most likely describe the amount of my participation in extracurricular activities as: superior above average average below average don't know
10	My high school classmates would most likely describe my leadership in extracurricular activities as: superior above average average below average don't know
11	The number of different high school sports I participated in was: 4 or more 3 2 1 didn't play sports
12	The number of letters I received in high school sports was: 4 or more 3 2 1 0
13	The number of high school clubs and organized activities (such as band, newspaper, etc.) in which I participated was: 4 or more 3 2 1 didn't participate
14	My final year in high school, I was absent: more than 15 days 10 to 14 days 5 to 9 days fewer than five days never

APPENDIX B
Applicant Background Assessment (ABA)

15	During my years in high school, I was singled out for disciplinary reasons: 5 or more times 3 or 4 times twice once never
16	My class standing in high school put me in the: top 10% top 33% top 50% top 90% did not graduate from high school
17	The high school grade I most often received was: A B C D or lower don't remember
18	The number of high school courses which I failed was: 5 or more 3 or 4 2 1 none
19	The high school English grade I most often received was: A B C D or lower don't remember or didn't take English
20	The high school math grade I most often received was: A B C D or lower don't remember or didn't take math
21	The high school science grade I most often received was: A B C D or lower don't remember or didn't take science
22	The high school subject in which I received my lowest grades was: science math English history/social sciences physical education

APPENDIX B
Applicant Background Assessment (ABA)

23	The number of elected offices I held in high school was: 5 or more 3 to 4 2 1 none

ACADEMIC EXPERIENCE: UNDERGRADUATE COLLEGE	
24	My highest education level is: no college 1 to 2 years of college or associate degree 3 to 4 years of college, no degree Bachelor's degree advanced degree

25	During college the number of times I made the Dean's List was: 5 or more times 3 to 4 times 1 to 2 times never didn't go to college
26*	Prior to accepting my first job in my present job series, I last attended college as a full-time student: did not attend college less than a year prior to accepting my first job in my present series one year prior to accepting my first job in my present series 2 to 3 years prior to accepting my first job in my present series over 3 years prior to accepting my first job in my present series
27	During my last year in college, my average number of hours of paid employment per week was: more than 20 hours 10 to 20 hours fewer than 10 hours none didn't go to college
28	The number of different undergraduate colleges I attended prior to graduation was: 4 or more 3 2 didn't change colleges didn't go to college
29	The number of times I changed my college major before I selected the one in which I graduated was: 3 times or more 2 times 1 time didn't change majors didn't go to college
30	My class standing in college put me in the: top 10% top 33% top 50% bottom 50% didn't go to college

APPENDIX B
Applicant Background Assessment (ABA)

31	The college grade I most often received was: A B C D or lower didn't go to college
32	On a 4 point scale where A=4, my grade point average the first two years of college was: I did not go to college or went less than two years less than 2.90 2.90 to 3.19 3.20 to 3.49 3.50 or higher
33	My grade point average after the first two years of college was: I did not go to college or went less than two years less than 2.90 2.90 to 3.19 3.20 to 3.49 3.50 or higher
34	My grade point average in my college major was: I did not go to college or went less than two years less than 2.90 2.90 to 3.19 3.20 to 3.49 3.50 or higher
35	My overall grade point average in college was: I did not go to college or went less than two years less than 2.90 2.90 to 3.19 3.20 to 3.49 3.50 or higher
36	Of the following, the college subject in which I received my lowest grades was: science English math history/political science didn't go to college
37	The number of college courses in which I received a failing grade was: 3 or more 2 1 none didn't go to college
38*	At the time I applied for my present job series, my undergraduate education consisted of having completed: less than 30 semester hours (45 quarter hours) 30 to 59 semester hours (45 to 89 quarter hours 60 to 90 semester hours (90 to 134 quarter hours) more than 90 semester hours (135 quarter hours) but no degree Bachelor's Degree

APPENDIX B
Applicant Background Assessment (ABA)

39*	At the time I applied for my present job series, my graduate education consisted of having completed: 0 to 5 graduate semester hours (0 to 8 quarter hours) 6 to 11 graduate semester hours (9 to 17 quarter hours) 12 to 23 graduate semester hours (18 to 35 quarter hours) 24 graduate semester hours or more (36 quarter hours) Master's Degree, Ph.D. Degree, or other graduate degree
40	The college English grade I most often received was: A B C D or lower didn't take English or didn't go to college
41	The college math grade I most often received was: A B C D or lower didn't take math or didn't
42	The college science grade I most often received was: A B C D or lower didn't take science or didn't go to college

43	The number of times I elected non-required college English courses was: 3 or more 2 1 never didn't go to college
44	The number of times I elected non-required college math courses was: 3 or more 2 1 never didn't go to college
45	The number of times I elected non-required college science courses was: 3 or more 2 1 never didn't go to college
46	The proportion of my college expenses that I earned was: more than 50% 25% to 50% some but less than 25% none didn't go to college
47	The amount of my college expenses covered by scholastic scholarships was: more than 50% 25% to 50% some but less than 25%

APPENDIX B
Applicant Background Assessment (ABA)

	none didn't go to college
48	The amount of my college expenses covered by athletic scholarships was: more than 50% 25% to 50% some but less than 25% none didn't go to college
50*	Prior to accepting my first job in my present job series, I had been out of college for: 5 or more years 3 to 4 years 1 to 2 years less than one year didn't go to college or didn't graduate
51	The number of college clubs and organized activities (band, newspaper, etc.) in which I participated was: 3 or more 2 1 didn't participate didn't go to college
52	The number of letters I received in college sports was: 3 or more 2 1 0 didn't go to college
53	The number of student offices to which I was elected in college was: 3 or more 2 1 0 didn't go to college
54	The number of national scholastic honor societies I belong to in college was: 3 or more 2 1 0 didn't go to college
WORK EXPERIENCE	
55*	In the three years prior to accepting my first job in my present job series, the number of different paying jobs I held for more than two weeks was: 7 or more 5 to 6 3 to 4 1 to 2 none

APPENDIX B
Applicant Background Assessment (ABA)

56*	In the three years immediately before accepting my first job in my present job series, the number of different full or part-time jobs I applied for was: none 1 to 2 3 to 4 5 to 6 7 or more
57*	Prior to accepting my first job in my present job series, I had been employed in work similar to that of my present job for: never employed in a similar job less than 1 year 1 to 2 years 3 to 4 years over 5 years
58*	In the three years before accepting my first job in my present job series, the number of promotions I received in all previous jobs was: not employed 0 1 2 3 or more
59*	I left my last full-time job (or job series) because: I was laid off or discharged there was little chance for advancement or increase in pay important personal reasons - such as moving or pregnancy something else have never had a full time job
60*	Prior to accepting my first job in my present job series, I worked on my last full-time job (or job series): have not held full-time job less then six months 6 months up to a year one to two years more than two years

61*	Prior to accepting my first job in my present job series, the number of different federal agencies I worked for (not :including military service) was: 0 1 2 3 4 or more
62*	I learned about the opportunity to apply for my present job series through: a public notice or media advertisement a friend or relative college recruitment working in some other capacity for the agency some other way
63	My military service was: none non-career enlisted non-career officer career enlisted career officer

APPENDIX B
Applicant Background Assessment (ABA)

64*	My employment status prior to accepting my first job in my present job series was: employed full-time employed part-time student, not employed self-employed unemployed
65*	The number of months I was unemployed during the three years immediately before accepting my first job in my present job series was: 0 1 to 2 3 to 4 5 to 6 7 or more
66*	Prior to accepting my first job in my present job series, I worked extra hours during evenings or on weekends: much more often than most persons in the job somewhat more often than most persons in the job about the same as most persons in the job somewhat less often than most persons in the job not employed prior to present job
67*	In the three years immediately before accepting my first job in my present job series, my work experience (military or civilian) was in: professional or administrative occupations clerical or sales occupations service occupations trades or labor occupations not employed during the three years immediately before accepting my present job
68*	On my last job (prior to accepting my first job in my present job series), my supervisor rated me as: outstanding above average average below average not employed or received no rating

69*	Prior to accepting my first job in my present job series, I was late (tardy for work): once or twice a year or less once or twice in a six month period once or twice a month once or twice a week not employed prior to present job
70*	In the three years prior to accepting my first job in my present job series, the number of formal awards I received for my job performance was: not employed prior to present job 0 1 2 3 or more
71	The amount of time I have been out of work between jobs usually has been: never out of work less than one month 1 to 2 months 3 to 4 months 5 or more months

APPENDIX B
Applicant Background Assessment (ABA)

72*	In the three years prior to accepting my first job in my present job series, the number of formal suggestions I submitted to my former employer(s) was: Not employed prior to present job 0 1 2 3 or more
73	The age at which I first started to earn money (other than an allowance) was: Less than 12 years old 12 to 13 years old 14 to 15 years old 16 to 17 years old 18 years or older
74*	In the year before accepting my first job in my present job series, the number of times I had been late for work (or class) was: More than 14 times 10 to 14 times 5 to 9 times fewer than five times never
75*	In the three years prior to accepting my first job in my present job series, the number of jobs I had been fired from was: 5 or more 3 to 4 2 1 none
76*	Prior to accepting my first job in my present job series, I was asked to serve as supervisor in my boss' absence: somewhat more often than most about the same as most others somewhat less often than most much less often than most not employed prior to present job

APPENDIX B
Applicant Background Assessment (ABA)

77*	Prior to accepting the first job in my present job series, I was selected to attend training: somewhat more often than most about the same as most others somewhat less often than most much less often than most not employed prior to present job
78*	Prior to accepting my first job in my present job series, I was chosen to serve on special task forces or committees at work: somewhat more often than most about the same as most others somewhat less often than most much less often than most not employed prior to present job

SKILLS	
79	The number of civic organizations or social organizations (which have regular meetings and a defined membership) that I belonged to prior to accepting my present job is: None 1 2 or 3 4 or 6 7 or more
80	Which one of the following have you ever organized or assisted in organizing? If you organized more than one, mark the one most important to you. Athletic team or sport competition Financial or charity campaign to raise funds Some other civic, social, work related, or professional organization Have never organized or assisted in organizing any club or group
81	The number of elective offices (other than in high school or college organizations) I have held in the last five years is: None 1 or 2 3 or 4 5 or 6 7 or more
82	In organizations to which I belong, my participation is best described as: do not belong to any organizations not very active a regular member but not an office holder have held at one important office have held several important offices
83	My previous supervisor (or teachers if not previously employed) would probably describe my attendance record as: more worse than my peers somewhat worse than my peers about the same as my peers somewhat better than my peers much better than my peers

APPENDIX B
Applicant Background Assessment (ABA)

84	My previous supervisor (or teachers if not previously employed) would most likely describe my problem solving skills as: superior above average average below average don't know

85	My previous supervisor (or teachers if not previously employed) would most likely describe my skill at thinking on my feet as: superior above average average below average don't know
86	My previous supervisor (or teachers if not previously employed) would likely describe the amount of supervision that I need as: more than average average less than average very little don't know
87	My previous supervisor (or teachers if not previously employed) would most likely describe my dependability as: superior above average average below average don't know
88	My previous supervisor (or teachers if not previously employed) would most likely describe the speed at which I work as: superior above average average below average don't know
89	My previous supervisor (or teachers if not previously employed) would most likely describe the amount of time I needed to complete assignments as: a great deal more than average average less than average don't know
90	My previous supervisor (or teachers if not previously employed) would most likely describe my skill at meeting deadlines under pressure as: superior above average average below average don't know

APPENDIX B
Applicant Background Assessment (ABA)

91	My previous supervisor (or teachers if not previously employed) would most likely describe me as taking on more than I can handle: Most of the time a great deal of the time sometimes infrequently don't know
92	My previous supervisor (or teachers if not previously employed) would most likely describe me as mastering my assignments: Most of the time a great deal of the time sometimes infrequently don't know

93	My previous supervisor (or teachers if not previously employed) would most likely describe my supervisory potential as: superior above average average below average don't know
94	My previous supervisor (or teachers if not previously employed) would most likely describe my skill at getting along with others as: superior above average average below average don't know
95	My previous supervisor (or teachers if not previously employed) would most likely describe my oral communication skills as: superior above average average below average don't know
96	My previous supervisor (or teachers if not previously employed) would most likely describe my self control as: superior above average average below average don't know
97	My previous supervisor (or teachers if not previously employed) would most likely describe my responsiveness to other person's viewpoints as: superior above average average below average don't know

APPENDIX B
Applicant Background Assessment (ABA)

98	My previous supervisor (or teachers if not previously employed) would most likely describe my skill at speaking before a group as: superior above average average below average don't know
99	My previous supervisor (or teachers if not previously employed) would most likely describe my logical reasoning skills as: superior above average average below average don't know
100	My previous supervisor (or teachers if not previously employed) would most likely describe my planning and organizing skills as: superior above average average below average don't know
101	My previous supervisor (or teachers if not previously employed) would most likely describe my analytical skills as: superior above average average below average don't know
102	My previous supervisor (or teachers if not previously employed) would most likely describe my basic math skills as: superior above average average below average don't know
103	My previous supervisor (or teachers if not previously employed) would most likely describe my vocabulary as: superior above average average below average don't know
104	My previous supervisor (or teachers if not previously employed) would most likely rate my writing skills as: superior above average average below average don't know

APPENDIX B
Applicant Background Assessment (ABA)

105	My previous supervisor (or teachers if not previously employed) would most likely rate my speed of reading skill as: superior above average average below average don't know
106	My previous supervisor (or teachers if not previously employed) would most likely rate my reading comprehension skill as: superior above average average below average don't know
107	My previous supervisor (or teachers if not previously employed) would most likely rate my skill at doing several different jobs at the same time as: superior above average average below average don't know
108	My previous supervisor (or teachers if not previously employed) would most likely describe my attention to detail as: superior above average average below average don't know
109	My previous supervisor (or teachers if not previously employed) would most likely describe my ability to recall facts and details of information as: superior above average average below average don't know
110	My previous supervisor (or teachers if not previously employed) would most likely describe my skill at getting work done on time as: superior above average average below average don't know
111	The number of years of leadership experience I have had (such as work supervisor, commissioned or non-commissioned officer, scout patrol leader, school or social club president, athletic captain, etc.) is: 5 or more years 3 or 4 years 2 years 1 year

APPENDIX B
Applicant Background Assessment (ABA)

112	In the past six months, the average number of hours per week I spent reading newspapers, books, magazines, etc. outside of work is: 5 or more hours per week 3 to 4 hours per week 2 hours per week 1 hour per week less than 1 hour per week
113	My peers would likely rate my interpersonal skills as: superior above average average below average don't know
114	On a list of 100 typical people in the kind of job I can do best, my peers would probably place me in the: top 10% top 25% top 50% top 75% top 90%
115	In terms of punctuality, my peers would probably say that I usually arrive: much later than most lather than most on time earlier than most much earlier than most
116	If you were to ask my peers, they would probably say that the amount of recognition I receive relative to my accomplishments is: a great deal less than deserved somewhat less than deserved as much as is deserved somewhat more than deserved much more than deserved

117	My peers would probably say that the highest level I could reach if I chose a career in a major corporation would be: a top level executive (e.g. vice president) a middle manager a first level supervisor a professional or technical expert other non-supervisory technical or administrative position
118	My peers would probably describe me as a person who: never takes chances hardly ever takes chances sometimes take chances often takes chances very often takes chances
119	My peers would probably describe me as: much more aggressive than most of my peers somewhat more aggressive than most of my peers about as aggressive as most of my peers somewhat less aggressive than most of my peers much less aggressive than most of my peers

APPENDIX B
Applicant Background Assessment (ABA)

120	My peers would probably say that getting me to change once I have made up my mind is: much harder than most somewhat harder than most about the same as most somewhat easier than most much easier than most
121	Which of the following communication situations would your peers say you would handle best? writing a lengthy report giving a lecture or speech to a large group mixing and conversing with a room full of strangers discussing a topic with another individual don't know
122	Which of the following would your peers say describes your behavior in a group situation? you freely express your views, and sway the group considerably you freely express your views, but the group does not always share them you are reluctant to express your views, but when you do they are usually well received you usually don't express your views don't know
123	Which of the following would your peers say describes your behavior in a social situation? always at ease in social situation almost always at ease in a social situation generally at ease in a social situation occasionally at ease in a social situation don't know
124	My peers would probably say that having someone criticize my performance (i.e. point out a mistake) bothers me: much less than most somewhat less than most about the same as most somewhat more than most much more than most
125	My peers would probably describe me as being: much more confident than most somewhat more confident than most about as confident as anyone else somewhat less confident than most much less confident than most

APPENDIX B
Applicant Background Assessment (ABA)

126	Which of the following would your peers consider your weakest trait? learning new things quickly composing effective written report working with and getting along with other people speaking and expressing yourself effectively to others working well under pressure
127	Which of the following would your peers consider your strongest trait? learning new things quickly composing effective written report working with and getting along with other people speaking and expressing yourself effectively to others working well under pressure
128	My peers would likely rate my skill in influencing people to my point of view as: superior above average average below average don't know
129	Compared to others in my unit, my rate of promotion in the military was: much faster than most somewhat faster than most about the same as most somewhat slower than most never served in the military
130	Compared to others on my last full-time job, my rate of promotion was: much faster than most somewhat faster than most about the same as most somewhat slower than most not employed full-time prior to present job
131	Prior to accepting my present job I: never worked for this agency worked part-time for this agency while in college worked for this agency during summer vacations while in college worked full-time for this agency for a period of but then resigned was employed full-time with this agency immediately prior to accepting my present job
132	Before I joined the government, the information I had about the type of work that air traffic controllers are expected to do was: none practically no information some information quite a bit knew in considerable detail
133*	Prior to accepting my first job in my present job series, the amount of formal training that I had (other than college) related directly to my present job was: less than 6 months 6 months to a year 1 to 2 years 3 to 4 years 5 or more years

APPENDIX B
Applicant Background Assessment (ABA)

134	During my teens, I usually spent most of my summers (choose one): taking life easy attending summer school attending honors classes working part-time working full-time

135	Before accepting my present job, the length of time I had worked shift work was: never worked shift work less than 6 months 6 to 12 months 13 months to 2 years more than 2 years
136	The number of times in the past five years I was denied an award I deserved is: never once or twice three or four times five or six times seven or more times
137	In the past year, I have been annoyed by my coworkers: never rarely occasionally frequently constantly
138	Compared to my peers, I find myself leading others: much more often than most somewhat more often than most about the same as most somewhat less than most much less often than most
139	Compared to my coworkers, people come to me for advice: much more often than most somewhat more often than most about the same as most somewhat less than most much less often than most
140	if I could have any full-time job I wanted, the reason I would pick the job which I would finally choose is that: I would be recognized for the work I do I would be with people I really like I would have the freedom to be creative I would have great possibilities for monetary rewards I could do the kind of work that I find very interesting
141	when I think about being an air traffic controller, the first thing that turns me off most about the job is that: achieving anything of significance might be difficult doing the same things over and over might be boring lacking control over my work activities would be frustrating having little prestige as a controller would be unsatisfying working under constant pressure could be very hard

APPENDIX B
Applicant Background Assessment (ABA)

142	The aspect of being an air traffic controller that appeals to me most is that: my job is secure in the future I'm responsible for the safety of many others I'll receive a good salary which will grow I'll be constantly challenged to resolve situations which arise the work will always be interesting

U.S. Department of Transportation
Federal Aviation Administration

Memorandum

Subject: FAA data for dissertation purposes Date: March 2, 2000

From: Manager, Training & Organizational Laboratory, AAM-520

To: William L. Farmer, AAM-520

This letter is to re-affirm that, in support of agency research objectives on alternative selection measures for the air traffic control specialist occupation, you are granted permission to use archival bio-demographic, cognitive aptitiude test, and training performance data and measures. Specifically, you are granted use of these data and measures for your dissertation on the characteristics of biodata keys as a function of scaling method, sample size, and criterion. The data are provided for research purposes only and may not be used for any commercial purpose. You agree to acknowledge the Federal Aviation Administration (FAA) as the source for your research data, andprovide a bound copy of your doctoral dissertaion to the FAA.

Edna R. Fiedler, Ph D
Manager, Training and Organizational Research Laboratory