This is to certify that the

thesis entitled

NOT ALL ABILITY DATA ARE THE SAME:
JOB CLUSTERING WITH GATB DATA

presented by

Patrick Daniel Converse

has been accepted towards fulfillment
of the requirements for

___M.A.___degree in __Psychology__

_____
Major professor

Date _August 13, 2002_

PLACE IN RETURN BOX to remove this checkout from your record.
TO AVOID FINES return on or before date due.
MAY BE RECALLED with earlier due date if requested.

DATE DUE	DATE DUE	DATE DUE

NOT ALL ABILITY DATA ARE THE SAME: JOB CLUSTERING WITH GATB
DATA

By

Patrick Daniel Converse

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF ARTS

Department of Psychology

2002

ABSTRACT

NOT ALL ABILITY DATA ARE THE SAME: JOB CLUSTERING WITH GATB
DATA

By

Patrick Daniel Converse

This study examines similarities and differences in occupational clusters resulting

from three types of occupational data, as well as the implications of these similarities and

differences for some of the purposes to which clusters are put in practice. Previous

research has indicated that occupational data from different psychological domains (e.g.,

abilities vs. tasks) can result in substantially different job clusters. This study extends

this previous work by examining the effects of different methods of developing

occupational data within a single domain (abilities) on job clusters. Specifically, the

same occupations were clustered using three different types of occupational aptitude

requirements data: mean job incumbent General Aptitude Test Battery (GATB) scores,

job analyst GATB ratings, and regression-estimated GATB scores. Implications of

similarities and differences in job clusters resulting from these three different types of

aptitude data were then examined for three purposes: selection test validation, job

evaluation, and career exploration using the Occupational Information Network

(O*NET). Results indicated that these three types of data produced substantially

different job clusters. However, these differences did not appear to have clear

implications for test validation, job evaluation, or career exploration using O*NET.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Job classification underlies numerous personnel-related activities. Rather than serving as an end in itself, classification is often used as a tool to assist other personnel-related functions (Pearlman, 1980). For example, clustering individual positions into jobs and/or clustering jobs into higher level groups (e.g., job families) plays a vital role in activities such as performance appraisal (e.g., Cornelius, Hakel, & Sackett, 1979), test validation (Arvey & Mossholder, 1977), job evaluation (Pearlman, 1980), career-path planning (Harvey, 1986), and vocational guidance (e.g., in the U.S. Department of Labor's O*NET or Occupational Information Network; Peterson, Mumford, Borman, Jeanneret, & Fleishman, 1999). In each of these cases, job clustering is similar to factor analysis in that it reduces a large number of jobs into smaller manageable groups to simplify and amplify relevant similarities and differences. For instance, rather than developing distinct performance appraisal instruments for each individual job in an organization, job clustering allows personnel psychologists to develop instruments for a smaller number of job families. What would be a cumbersome, costly, and time-consuming task becomes a more manageable task that is less expensive and less time consuming, yet hopefully just as useful. The total number of appraisal instruments can be reduced to a manageable and appropriate size, assuming that (1) individual positions - and individual jobs - can be aggregated into job families on relevant cross-job characteristics, and (2) sacrificing unique information about individual positions and individual jobs does not materially affect the goals of performance appraisal.

Profiles

For the most part, clusters of jobs are assumed to have similar *profiles* of cross-job characteristics. An ability profile is a set of scores (e.g., verbal and math scores) used to describe an individual or occupation. Each score in the set represents the person or job's standing on a different variable. Figure 1 presents an example of a profile. In this case, an occupation is described by its ability requirements for eight aptitudes (each aptitude is rated on a scale from 1 to 5). Through quantitative or rational methods, profiles such as this can be used to cluster or categorize jobs according to their standing on multiple variables.

Present Study

The present study examines how job clusters differ depending upon what type of data is used to profile jobs. Specifically, three different types of occupational aptitude requirement profiles are used separately to cluster the same set of jobs: job analyst ratings, mean incumbent test scores, and regression-estimated scores. Results will highlight similarities and differences in the cluster solutions resulting from these three types of data, as well as the implications of these similarities and differences for some of the purposes to which clusters are put in practice.

As noted above, job clustering underlies several personnel-related functions. Because it often plays such an integral part in these activities, job clustering can have an important influence on their effectiveness. For instance, a particular clustering method or type of occupational data may tend to produce clusters that are inappropriate or ineffective for a given personnel function, whereas another method or data type might tend to produce more useful clusters for that purpose. For example, a clustering method

or data type that tends to produce job clusters with a substantial amount of within-cluster variability in pay rates would not be useful for job evaluation purposes (the process through which occupational pay levels are determined), suggesting that this method or data type should not be used for these situations. However, a method or data type that tends to produce clusters with little within-cluster variability in pay rates would likely be more useful for this purpose, suggesting that the method or data type should be used for these situations. Thus, it is important to uncover the strengths and weakness of clustering techniques, both in terms of clustering methods and the type of data used, for various functions in order to identify techniques that are most likely to be useful in a given situation.

Although a substantial amount of research has examined the effectiveness of various quantitative clustering methods (e.g., Milligan, 1981; Milligan & Cooper, 1987; Colihan & Burger, 1995), less research has focused on the type of data used for clustering. In addition, the limited research examining issues of data type has focused only on clusters resulting from data from different psychological domains (e.g., clusters resulting from ability data versus clusters resulting from task data), and has not considered the potential effects of different methods of developing job characteristic data within a single domain (e.g., abilities) on job clusters. This paper examines some of these effects.

Specifically, because it is often difficult to identify the strengths and weaknesses of clustering techniques a priori due to the complexity of clustering and its influence on personnel functions, this study takes an empirical approach to examine the strengths and weaknesses of different types of occupational aptitude data for clustering for a few of the

purposes to which clusters are put. Clusters resulting from three types of aptitude data (job analyst ratings, mean incumbent test scores, and regression-estimated scores) that differ only in the method used to generate them (not in the psychological domain to which they belong) are examined in terms of their effectiveness for three purposes: test validation, job evaluation, and career exploration.

Note that this study focuses only on the ability-requirement characteristics of occupations. It may be reasonable to use other job characteristic domains such as interests and personality to cluster occupations for several purposes (e.g., educational and vocational guidance). Examining domains beyond ability, either separately or in combination with ability, is beyond the scope of this study but could be a reasonable extension of the present work. The purpose of this study is to examine different types of data within the ability domain, as abilities are also used to cluster jobs for a variety of purposes (e.g., test validation, vocational and educational guidance, job placement) and are important determinants of job performance, person-job fit, and other educational, work, and career outcomes.

The remainder of this introduction is organized into three major sections. First, previous research examining the effects of different types of occupational data on job clustering is described. Next, the three types of aptitude data used in this study are described in terms of their development (the source(s) from which these data stem and the processes involved in their generation), the constructs they measure, and potential bias/inaccuracy that may be present in each type. Finally, implications of similarities and differences in these types of data and the clusters they produce are discussed for three personnel-related functions: test validation, job evaluation, and career exploration.

Previous Research

Background

Previous research examining the effect of different types of data on job clustering

has tended to focus on data across psychological domains. Rather than examining

different methods of developing job characteristic data within a single domain (e.g.,

abilities), these studies have focused on data from different domains (e.g., abilities versus

tasks). In general, this research indicates that job characteristic data from different

psychological domains can result in substantially different job clusters (e.g., Ghiselli,

1966; Pearlman, 1980; Cornelius, Carron, & Collins, 1979). The implication of this

finding is that the objective of classification should, first and foremost, determine which

of these domains is relevant for developing job characteristic profiles. If cluster solutions

differ depending upon the type of data used, then type of data should be chosen carefully,

according to the objective of clustering.

Hartman, Mumford, and Mueller (1992) reported one exception to this general

finding. Hartman, et al. compared job clusters resulting from data reflecting the types of

tasks performed to those resulting from data reflecting the knowledge, skills, and abilities

(KSAs) needed to perform the job. They found that more than half of the jobs were

placed in the same family across the two data types, concluding that the job

classifications "displayed some generalizability across different measurement formats"

(Hartman, et al., 1992, p. 208).

However, regardless of outcome, research in this area seems to have had a couple

of limitations. First, when examining different types of data, some studies have

confounded psychological domains with the methods used to develop the data. That is,

5

the types of data used to cluster jobs in these studies differ in terms of both the psychological domain to which they belong and the manner in which they were developed, making it difficult to interpret similarities and differences in the cluster solutions generated in these studies. For instance, Ghiselli (1966) found that rational groupings of jobs based on similar work were quite different from groupings based on similar validity patterns for intellectual-perceptual, spatial-mechanical, and motor ability tests. Ghiselli first developed job clusters by rationally grouping jobs according to the nature of the work performed in these jobs: Managerial, Clerical, Sales, Protective, Service, Vehicle, Trades and Crafts, and Industrial Occupations. He then clustered these same jobs according to their patterns of criterion-related validity coefficients for intellectual-perceptual, spatial-mechanical, and motor ability tests, finding that these two types of data led to dissimilar job groupings. Ghiselli noted that the differences between the two sets of job groupings were not systematic in obvious ways. He observed only that the clusters resulting from the ability data were made up of jobs with little apparent similarity, whereas clusters stemming from the nature of the work performed were made up of jobs that had obvious similarities but often differed substantially in terms of their patterns of criterion-related validity.

Note, however, that these two cluster solutions are based on data that differ both in the psychological domain to which they belong and the manner in which they were developed. On the one hand, rational job clusters were based on subjective impressions of the type of work performed. Thus, information developed through subjective impressions and belonging to a task domain produced these clusters. On the other hand, a second set of job clusters was based on patterns of criterion-related validity coefficients

for ability tests. Thus, data developed by correlating test scores and job performance measures and belonging to the abilities domain produced these clusters. Therefore, it is difficult to determine whether the two types of data produced different job clusters due to differences in the psychological domains to which they belong or differences in the manner in which they were developed.

Second, even when research has avoided this limitation, it has tended to focus only on job clusters resulting from data stemming from different psychological domains, rather than clusters resulting from data produced by different methods. For example, Cornelius et al. (1979) examined how data from different psychological domains might affect jobs clusters with a small sample of seven foreman jobs using task statement ratings, Position Analysis Questionnaire (PAQ) dimensions, and ability rating data. Data analysis on each type of data resulted in different job clusters even when applying the same hierarchical clustering method (Ward's minimum variance technique) to the same seven jobs each time. Task statement data resulted in three or five clusters, depending upon which criterion of task overlap was adopted. By contrast, PAQ data clustering resulted in only one cluster containing all seven jobs, and ability-rating data resulted in three clusters that differed from the task statement clusters. As the authors mention, the differences between these three cluster solutions seem to make sense in light of the differences between the types of data used to produce the clusters. For instance, PAQ data may not have resulted in distinct job clusters because the seven jobs were all foreman jobs, and the PAQ was designed for application (and for making distinctions) across a wide variety of jobs. Therefore, differences between the seven foremen jobs may have been too subtle to be detected by PAQ data. In contrast, task data seem to have

resulted in finer distinctions between jobs than the ability data. This may simply indicate that analyzed jobs differed more in terms of the actual tasks performed than in terms of the underlying abilities needed to perform these tasks.

Thus, this study indicates that data from different domains, developed in generally the same manner (through analyst ratings), tend to produce very different job clusters. Again, however, this research does not address the issue of the effect of different methods of developing job characteristic data on job clusters. The present study addresses this issue.

Same Jobs, Same Construct Domains

Studies have shown that job descriptor profiles from different broad psychological domains result in different job groupings, but little research attention has focused on how, even from within the same psychological domain, job groupings are affected by different types of profile data (e.g., test-score data, field analyst job analysis data). Many different types of profile data can be developed within the same psychological domain. These types of data can differ both in the specific constructs measured within the broad construct domain and in how the constructs are measured. For example, within the abilities domain, different constructs such as verbal ability, numerical ability, or motor coordination can be measured, and the same constructs can be measured in different ways, such as by ability tests, behavioral samples, or expert ratings. From the research cited previously, we might expect job profile data that include different constructs to yield different job groupings even though these constructs are from the same broad psychological domain (e.g., Pearlman, 1980; Cornelius et al., 1979).

8

However, less is known about the extent to which job profile data consisting of the same constructs, but differing in how these constructs are measured, might yield different job groupings. For example, the same jobs may be clustered differently when using average incumbent performance on numerical ability tests versus job analyst ratings of the necessary levels of these same numerical abilities. The recent development of the Occupational Information Network (O*NET) Career Explorer tools (as described by McCloy, Campbell, & Oswald, 1999) provides an opportunity to examine this latter effect. O*NET is the Department of Labor's computerized occupational information tool developed to replace and extend the *Dictionary of Occupational Titles* (DOT). The O*NET database is organized around an overarching 'Content Model' that encompasses theories across major psychological domains, covering individual differences as well as psychological and situational characteristics about work and the worker (Dye & Silver, 1999).

Among other resources, O*NET Career Explorer includes the "Ability Profiler," which helps individuals just entering careers or in mid-career transition focus their career-search activities. The Ability Profiler uses General Aptitude Test Battery (GATB; U.S. Department of Labor, 1979) subtests to measure clients' ability levels on up to nine aptitudes (Verbal Ability, Arithmetic Reasoning, Computation, Spatial Ability, Form Perception, Clerical Perception, Motor Coordination, Finger Dexterity, and Manual Dexterity).[1] The Profiler then compares individuals' ability profiles with ability profiles for 1,172 job clusters or Occupational Units (OUs), presenting the client with a subset of OUs that most closely fits his/her profile. The Profiler defines *fit* as a correspondence between client and OU ability profile shape represented by the correlation between

9

profiles. The highest 50 correlations between client score profiles and each OU score profile are presented (McCloy, et al., 1999). Clients can then narrow their searches within the subset of OUs that fits their ability profile, considering information such as specialized skill and training requirements, salary offerings, and future hiring prospects for jobs within OUs.

Note that although OUs themselves often comprise DOT job groups, each OU still only comprises a relatively limited set of occupations. Thus, in order to give clients a wider range of jobs for career exploration, the Ability Profiler presents clients with several OUs that match their ability profile.

It should also be noted that although the original version of O*NET included job classifications known as Occupational Units, the O*NET has been recently updated to adapt to the more widely-accepted Standard Occupational Classification (SOC) System. The SOC was developed by the U.S. Department of Labor to be a universal occupational classification system and, by law, is now used by all federal agencies collecting occupational information (Bureau of Labor Statistics, 2000). OUs are based on the Occupational Employment Statistics (OES) classification system and were created by clustering DOT occupations (U.S. Department of Labor, 1998). An OU-SOC "crosswalk" that fits OUs into the SOC system (National Crosswalk Service Center, 2001) was used to make the transition to this newer classification scheme.

O*NET's Ability Profiler was developed using three types of ability data: field-analyst-rated GATB profiles, actual GATB test score profiles, and regression-estimated GATB profiles. The test-score-based and field-analyst-rated profiles were used to develop regression-estimated profiles for all 12,000+ DOT occupations. Regression-

estimated profiles were then aggregated to the OU level. These aggregated profiles

constitute the occupational ability profiles that the Ability Profiler compares with client

profiles to generate job suggestions.

The present study uses these three types of data - field-analyst-rated, actual, and

regression-estimated GATB profiles - to examine how different types of profile data from

the ability domain affect job groupings. That is, although all three types of data represent

measurements of the same ability constructs (the nine GATB aptitudes), the data come

from very different sources, and therefore the processes generating these data are also

different. These differences may have resulted in substantively different ability profiles

among the three types of data for the same jobs, which in turn may lead to differing job

groupings.

Job Classification According to Ability Profiles

As discussed, numerous types of occupational data (e.g., tasks, interests, abilities)

can be used to cluster or classify occupations for various reasons (e.g., the development

of performance appraisal instruments, vocational guidance, test validation). Ability data

in particular have been used to cluster occupations for some of these purposes (e.g.,

vocational counseling; Gottfredson, 1986). In many cases, ability-based occupational

classifications are based on overall ability requirements or level of job complexity.

Several researchers (e.g., Spaeth, 1979; Gottfredson, 1986; Desmarais & Sackett, 1993)

have attempted to differentiate occupations according general cognitive ability

requirements or overall job complexity.

Some of these researchers have also suggested and attempted to provide evidence

that occupations can be further classified according to their requirements for specific

abilities or the shape of their cognitive ability requirement profiles (Gottfredson, 1986; Desmarais & Sackett, 1993). In other words, these researchers argue that occupations can be differentiated or classified not only according to their general ability requirements, but also according to their patterns of requirements across different types of abilities. For instance, in the development of an occupational classification system called the Occupational Aptitude Patterns (OAP) Map, Gottfredson (1986) placed occupations into clusters according to overall ability requirements, arguing that general cognitive ability requirements are "the single most important aptitude distinction among jobs" (p. 285). However, she also proposed that within these occupational levels, different types of occupations differ in the shape of their cognitive ability requirement profiles, but was unable to present much evidence for this.

Desmarais and Sackett (1993) examined the validity of the OAP Map by placing the positions held by a nationally representative sample of employees into the classification system. In general, this study supported the OAP Map structure. In addition, the researchers found some evidence that occupations can be differentiated according to specific abilities after the effects of general cognitive ability requirements are taken into account. For instance, individuals in Bureaucratic or Social jobs (two of the four general occupational fields included in the OAP Map) tended to score well on a speededness variable and poorly on a scientific/mechanical ability variable, whereas individuals in Physical jobs tended to demonstrate the opposite pattern. Thus, this study indicates it may be possible to classify occupations into broad categories according to their patterns of ability requirements.

The present study classifies occupations according to their patterns of ability requirements using three types of data (field-analyst-rated, actual, and regression-estimated GATB profiles) to examine similarities and differences in occupational clusters resulting from these different data types as well as the implications of these similarities and differences for a few personnel-related activities. The following sections describe each of the three types of data.

Three Types of Aptitude Data

Job Analyst Ratings

Field-analyst profiles were collected during the DOT's development. Since its third edition (published in 1965), the DOT has included job analysts' ratings of several important worker traits such as aptitudes, temperaments, and interests (Miller, Treiman, Cain, & Roos, 1980). In order to keep up with occupational changes, these ratings have been verified, revised, or added for each edition since the third (i.e., for the 1977 fourth edition, the 1982 supplement, the 1986 supplement, and the 1991 revised fourth edition; U.S. Department of Labor, 1991). This study uses ratings from the 1991 revised fourth edition, the most recent edition of the DOT, although new rating data are currently being collected for the O*NET.

To develop occupational aptitude profiles, expert job analysts first observed individual jobs and wrote descriptions of their purposes and tasks. On the basis of these descriptions and other observations, analysts then rated each occupation on 11 aptitudes: the nine GATB aptitudes, plus Eye-Hand-Foot Coordination and Color Discrimination. For each job rated, analysts estimated on a 1-5 scale the level of each aptitude required of the worker for "average, satisfactory performance": from 1 = extremely high aptitude

ability (top 10%) to 5 = markedly low aptitude ability (bottom 10%; U.S. Department of Labor, 1991, p. 9-2). Aptitude profiles from similar jobs were then aggregated to the DOT-occupation level such that each DOT occupation's rating on each of the 11 aptitudes reflects the modal value of the ratings from its constituent jobs (Cain & Green, 1983).

Actual GATB Test Scores

Actual GATB profiles were obtained from the test scores of workers with each of the 9 GATB aptitudes – General Intelligence (G), Verbal Ability (V), Numerical Ability (N), Spatial Ability (S), Form Perception (P), Clerical Ability (Q), Motor Coordination (K), Finger Dexterity (F), and Manual Dexterity (M) (cf. McCloy, et al., 1999). Averaged ability test scores result in a profile of the average abilities needed to perform a job satisfactorily.

Average incumbent test scores are assumed to reflect the ability levels required for average, satisfactory performance based on evidence indicating that over time individuals tend to gravitate toward jobs that are commensurate with their ability levels (e.g., Wilk, Desmarais, & Sackett, 1995; Wilk & Sackett, 1996). This research appears to indicate that attrition is likely at the high and low ends of the ability continuum, relative to the job's ability requirements, leaving individuals with appropriate ability levels who are likely to perform satisfactorily. Average test scores of these individuals will then reflect the average ability level of satisfactorily performing incumbents. Data for actual GATB ability profiles exist for 545 jobs where workers were tested with the GATB.

During the development of the Ability Profiler, McCloy, et al. (1999) generated ability profiles for all OUs, but because only 545 ability profiles existed at the DOT level, OU ability profiles not represented by these jobs had to be estimated from actual data. This was accomplished in two stages. The first stage involved generating ability score profiles for each DOT occupation. This required two steps. First, 48 predictor variables were used, constituting DOT job analysis information such as job analysts' ratings of such variables as Data, People, Things, and Specific Vocational Preparation – see Table 1. These variables were reduced via principal components analysis down to a set of seven promax-rotated component scores. Then, occupations' mean-ability scores were regressed on these component scores. This resulted in a set of regression weights that could then be used for any DOT occupation to predict its mean ability score. Again, this was done because actual GATB profiles existed for only 545 out of the 12,000+ DOT level occupations, yet the desire (by O*NET's Ability Profiler developers) was to have ability profiles for all 12,000+ DOT occupations.

The second stage involved computing ability score profiles for the OUs from the ability score profiles for the constituent DOT occupations. This was accomplished by computing the mean for each ability across all the constituent DOT occupations (for OUs with fewer than 7 or more than 300 DOT occupations), or across only the constituent DOT occupations with high loadings on the first principal component, as determined by principal components analysis (for OUs with more than 7 but less than 300 DOT occupations; see McCloy, et al., 1999 for further details on the development of the Ability Profiler's occupational ability profiles).

Thus, the O*NET Ability Profiler's development involved three distinct types of ability profiles: actual test score, regression-estimated, and analyst profiles. Although the three types of profiles are intended to measure the same aptitudes, the different processes through which each type of profile was developed may have resulted in different profiles among the three types of data, even for the same job. Average test score data were developed through incumbent ability testing, analyst data were developed through a cognitive estimation process, and regression data were developed through statistical prediction. Human judgments - even expert judgments - of necessary or minimally required abilities may not correspond with average incumbent performance on ability tests, which in turn may not correspond with regression-predicted ability estimates. Therefore, in many cases these distinct methods will produce different job profiles, which in turn will likely result in differing job clusters.

In general, analysts, incumbents, and regression equations produce profiles that may differ in terms of their level or elevation and/or their shape. For example, analysts' estimates of overall required ability levels may be lower than the level of actual aptitude scores, producing a difference in profile level across data types. Additionally, for a given job, analysts estimates of some ability levels may be lower than the level of actual aptitude scores for these abilities, while their estimates of other abilities may exceed aptitude scores, leading to a divergence in profile shape among data types. Both situations may have implications for job clustering based on these types of data.

The following sections describe both the *source* and *process* involved in generating each type of data. *Source* refers to the primary source of information from which each type of data was generated. For example, trained job analysts are the

information source involved in generating analyst-rated data, whereas job incumbents are the information source involved in generating test score data. *Process* refers to the series of steps leading up to generating each type of data as suggested by theory and the extant research literature. For example, analyst data were developed through a process of cognitive estimation in which job analysts were required to observe, encode, store, retrieve, and integrate job relevant information, whereas test data were developed through incumbent testing in which multiple job incumbents took a battery of tests and their scores were averaged to yield job-level estimates of aptitude requirements.

These source and process sections contain broad theoretical assertions and discussion that serve to help us understand better the similarities and differences behind these three types of data that may have produced similarities and differences in: (1) the actual constructs measured across data types, and (2) the biases or inaccuracies in measurement present among the types of data. Although all three types of data are measurements of required aptitudes, this does not guarantee that they measure precisely the same constructs, or that they are subject to the same biases/inaccuracies. It should be noted that many of the propositions in the following discussion of the constructs and potential biases/inaccuracies for each type of data are theoretical and cannot be evaluated or tested directly but can be supported by previous research.

Job Analyst Estimation of Ability Requirements

Information Source

Essentially, ratings of job requirements can come from three sources: job incumbents (those who actually perform the job), other organizational members (those who perform other jobs but have the opportunity to observe the job of interest – such as

supervisors, peers, etc.), and/or outside job analyst experts (those who do not regularly observe the job of interest, but rather are employed specifically to develop job characteristic data across occupations). The ability requirement ratings used in this study (those developed for the DOT) stem from the third source: job analysts. Trained job analysts were employed to develop job descriptions and ratings for each DOT occupation.

Intuitively, outside job analysts seem to be the most appropriate source for ability requirement ratings because analysts seem likely to produce the most accurate requirement ratings. Relative to outside analysts, incumbents and other organizational members appear to be at a disadvantage in terms of their ability to produce accurate ability ratings. For instance, considering the ego-involving nature of rating the ability requirements of one's own job, incumbents might be expected to tend to overestimate the levels of abilities required for their jobs.

In addition, incumbents, supervisors, and other organizational members are likely to have much less training and experience in job analysis than those specifically employed for job analysis efforts. Job analysts are employed expressly to rate numerous jobs across fields (e.g., occupations in engineering, sales, manufacturing), organizations (e.g., government organizations, private corporations), and job levels (e.g., entry level, middle-level managers, executives). They are more likely to have the 'across-jobs' perspective and knowledge necessary for developing accurate ability requirement ratings for analyzing abilities across jobs.

Incumbents and other organizational members, by contrast, are not likely to have the same breadth of knowledge concerning occupations' ability requirements relative to

18

other occupations or even other organizational contexts and thus are probably not as able to evaluate accordingly. Their relatively narrow perspective may then lead to less informative ability ratings across jobs because, in many cases - including the DOT data used in this study - ratings of the ability requirements of jobs need to be relative or norm-referenced ratings. That is, concrete criteria can inform but do not determine the levels of abilities required for jobs. Rather, the required ability levels for a particular job are determined by comparing the ability requirements of that job to those of other jobs. An occupation is rated as requiring a high level of ability if it requires more of this ability than most other jobs, and it is rated as requiring a low level of ability if it requires less of this ability than most other jobs. Thus, to the extent that raters lack the relevant cross-job knowledge and a cross-job perspective, the accuracy of their ratings are likely to suffer. Again, this is another reason for arguing that outside analysts may be the most appropriate source for ability requirement ratings.

However, evidence relevant to this issue does not appear to be completely supportive of this reasoning. Although much of the research comparing sources of job analysis data seems to have focused on incumbents and supervisors (e.g., Huber, 1991; Waldman, Yammarino, & Avolio, 1990; O'Reilly, 1973), some research has compared job analysis data obtained from analysts to those obtained from incumbents and other organizational members. This research indicates that data obtained from these different groups may not differ greatly. For instance, Smith and Hakel (1979) found that little difference existed between job incumbents, supervisors, job analysts, and a comparison group of college students in terms of their ability to analyze reliably a job using the Position Analysis Questionnaire (PAQ). The lowest correlation among the mean ratings

19

across all PAQ items for these judge categories was .89, and the data obtained from each of these groups were significant predictors of present pay levels, with uncorrected correlations between actual salary and predicted salary ranging from .39 for student raters to .67 for analysts. These finding led the authors to conclude that "who furnishes responses to a job analysis inventory makes little practical difference" (p. 677).

Studies by Fischer and Sobkow (1979) and Desmond and Weiss (1975) each obtained similar results. In these studies incumbents were asked to rate the ability requirements of their jobs (on the GATB aptitude dimensions using a 1-6 scale). Ratings were then compared to expert job analysts' ratings from the DOT. Results from both studies indicated that incumbents were able to produce reasonably reliable ratings, and the occupational ability patterns (OAPs), consisting of worker ratings of GATB abilities categorized as 'important' or 'not important' for each job, compared favorably with those derived from expert ratings. Workers and analysts produced similar patterns of 'important' and 'not important' abilities for each job. In addition, Desmond and Weiss (1975) found that OAPs derived from worker ratings compared favorably with those derived from supervisor ratings, as determined by a subjective evaluation of similarity in OAP patterns. Finally, in another study conducted by Desmond and Weiss (1973), supervisor ratings of the ability requirements of their subordinates' jobs were similar to expert ratings from the DOT. Once again, it appears that job analysts do not rate job requirements substantially different from incumbents and other organizational members.

It appears that these findings should be interpreted with some caution, however. These studies examined sources of ratings by comparing OAPs derived from incumbent, supervisor, and analyst ratings rather than by comparing the ratings themselves.

Although OAPs may be practically useful and converge across different types of raters, OAPs are relatively crude indicators of ability requirements, consisting only of dichotomous important/not important distinctions rather than actual ratings (e.g., on a 1-6 scale). The research under discussion does not address the fact that incumbents, other organizational members, and expert analysts may not be equally capable of providing more precise ratings of ability levels (e.g., ratings on a 1-6 scale).

Process

The process of rating jobs' ability requirements is one of cognitive judgment or estimation. As such, this process involves five basic aspects: observing behavior, encoding information about behavior, storing information, retrieving information, and integrating information (see Murphy & Cleveland, 1995, for a discussion of these processes in the context of performance appraisal). Ability requirements cannot be observed directly and thus must be inferred on the basis of other information. Thus, analysts estimate abilities based on job-relevant information usually gathered through some combination of short-term observation of incumbents at work, interviews with incumbents, supervisors, or peers, and job descriptions or other informational materials pertaining to the job of interest (e.g., training manuals). Job analysts then encode this information regarding what is done on the job and how tasks are carried out. Information is then stored in analysts' long-term memories, or stored externally (e.g., written down). For example, DOT rating information was often documented on paper for later use. At the time abilities are estimated, undocumented information must then be retrieved from long-term memory. Finally, analysts must integrate all of the job-relevant information, and make inferences and decisions about ability requirements based on this integration.

That is, because raters cannot observe ability requirements directly, they must observe

tasks and take in other job information (e.g., formal job requirements, training manuals,

the products of work) and make inferences regarding levels of required abilities. Thus,

the process of rating job ability requirements appears to involve two components: a *basic*

information processing component (observing, encoding, storing, retrieving, and

integrating information) as well as an *inference generating component* in which raters

estimate required ability levels based on the basic information process that informs their

understanding of the ability constructs, performance on tasks and the job, and in what

ways abilities are critical determinants of performance on the job.

Constructs. Analyst ratings are intended to measure the level of aptitudes

required for "average, satisfactory performance" on the job (presumably under normal

working conditions) (U.S. Department of Labor, 1991, p. 9-2). Thus, jobs' general

aptitude requirements for average performance can be considered the measured

constructs.

Bias/Inaccuracy. Bias or inaccuracy in estimating aptitude requirements may be

introduced in any of the six processes discussed above (observing, encoding, storing,

retrieving, integrating, and inference generating). Each of these processes may introduce

both systematic and unsystematic error into the final estimate. For example, research has

indicated that, during observation, unexpected characteristics (which are more salient)

tend to result in controlled processing of information, whereas behavior consistent with

one's expectations or stereotypes about the job will be noted and stored automatically

(Murphy & Cleveland, 1995; Feldman, 1981). This may mean that the unexpected

features of an occupation will tend to be consciously observed more frequently than

typical or expected features, and thus may influence aptitude ratings more regardless of the importance of these factors to aptitude ratings.

In addition, the encoding of job-relevant information might introduce error because this process may involve a complex process of categorization (Murphy & Cleveland, 1995). Specifically, rather than encoding specific information about each job, analysts may simply represent jobs as belonging to a category, based on its similarity to the prototype of that category. In general, it is likely that individuals then remember the category rather than the stimulus (Murphy & Cleveland, 1995). Thus, to the extent that the occupation does not fit the category, later ratings will be inaccurate, not just in terms of forgetting information but also recognizing information that was never in fact presented (Feldman, 1981).

The storage and retrieval processes may also introduce error in several ways. One model describing how these processes function is the *storage model* of memory, in which memory is conceptualized as consisting of several bins, each of which is designed to hold different types of information (Murphy & Cleveland, 1995). Generally, errors in the rating process would occur when the correct bins are not searched. In addition, the categorization described previously tends to influence and bias recall selectively, preventing contradictory evidence from appearing and eliciting confirming evidence (Feldman, 1981). Finally, recording information such as taking notes about the job during a job analysis may help in preserving specific details of what has been observed, but it also limits the observer to processing information in serial order. Thus, the analyst may not be able encode information about other stimuli present at the same time.

Recording information may help with some details - especially for long-term recall - at the expense of other details.

The integration and inference-generating processes will then produce inaccurate ratings to the extent that the information stored and retrieved is inaccurate, biased, or incomplete. These processes are completely dependent on the information available (and not available) to the analyst (Feldman, 1981). Thus, any error introduced due to the processes described above will influence integration and inference.

Morgeson and Campion's (1997) discussion of more general sources of potential inaccuracy in job analysis complements this discussion of potential error in the rating process. Their review overlaps some with the previous discussion, but takes a broader view of potential sources of error. Specifically, these authors discuss two broad categories of sources of potential inaccuracy in job analysis - social processes and cognitive processes - and 16 specific sources within these two categories (see Table 2).

It is not clear that the social influences on job analysis data discussed by Morgeson and Campion (1997) would have played much of a role in the development of DOT ability ratings. First, individual analysts, not groups of analysts, appear to have rated each job. Thus, social influence processes such as extremity shift, in which group members' opinions shift to more extreme judgments following group discussion, would not have been present. Second, because analysts, not incumbents, developed the ratings, self-presentation processes were also probably not influential. For example, it seems unlikely that in rating the ability requirements of others' jobs, analysts would have been influenced by impression management, or the desire to cast themselves in a favorable

light. Thus, these social processes probably did not have much of an influence on the rating process.

Whereas the social influences do not likely play much of a role in the DOT rating process, the cognitive sources of inaccuracy discussed by Morgeson and Campion could very well have influence. DOT analysts would have been subject to the same information processing limitations and biases that can be present in any subjective judgment task. For example, inadequate information may have biased the rating process, as analysts had only a limited amount of information about each occupation. In addition, order and contrast effects might have been present, where, for example, ratings of one job were influenced by the characteristics of jobs rated just prior to that job (see Morgeson & Campion, 1997 for examples in the selection interview and performance appraisal literature). It appears that any of the cognitive sources were potentially involved in the rating of jobs but it is impossible to say which of the sources were involved in the rating of any particular job. It is likely that DOT ability ratings were not immune to cognitive sources of inaccuracy in the rating process.

Additionally, some of the specific conditions under which DOT ratings were obtained may have produced a common type of rating error referred to as *halo error*. Halo error (or illusory halo) occurs when a general impression seeps into the ratings of individual categories, artificially inflating relationships among dimensions (Cooper, 1981). This is contrasted with true halo, or the extent to which dimensions are correlated in reality (Cooper, 1981; Murphy, Jako, & Anhalt, 1993). Thus, although it is likely that some true halo exists among occupational ability requirement dimensions (i.e., these

dimensions are correlated in reality), the concern here is that analyst ratings may reflect illusory halo (i.e., the relationships among these dimensions are artificially inflated).

In a critical review of the DOT, Miller, Treiman, Cain, and Roos (1980) reported several difficulties analysts had in using some DOT rating scales, such as ambiguous rating dimensions and inadequate instructions for analyzing jobs. As Gottfredson (1986) notes, these difficulties suggest that DOT ratings may have been obtained under conditions that often produce illusory halo. For example, Cooper (1981) identified six sources of halo in rating: undersampling (rater's insufficient sampling of ratee behavior), engulfing (ratings are colored by an overall impression or salient features), insufficient concreteness (rating categories are too abstract), insufficient rater motivation and knowledge, cognitive distortions (stored observations are distorted and information is lost and added), and correlated true score (categories are correlated, where some halo is true, not illusory).

Although all of these sources may have been present in the development of DOT ratings, undersampling, engulfing, and insufficient concreteness are particularly likely to have been influential given the conditions under which ratings were obtained. As noted by Miller, et al. (1980), analysts usually observed only one or two workers for each job and usually had to work rapidly so as not to disrupt the company's work schedule. These conditions likely led to undersampling and engulfing, where analysts were forced to base ratings primarily on overall impressions because incumbents' work behavior was not sampled adequately. Furthermore, analysts themselves reported difficulty in assigning scores in some cases due to ambiguity of the rating dimensions and inadequacy of rating

instructions (Miller, et al., 1980). This situation appears to correspond to what Cooper (1981) referred to as *insufficient concreteness*, another source of halo.

Finally, the categorization processes involved in encoding and storage of information by DOT analysts might also contribute to halo. As discussed previously, when encoding and storing occupational information, rather than encoding specific information about each job, analysts may represent occupations as belonging to a category, based on its similarity to the prototype of that category. Analysts are then more likely to remember the category rather than specifics about the jobs themselves. As a result, analysts may produce occupational aptitude profiles reflecting the general requirements of a category of jobs that are similar to the job of interest (i.e., the requirements associated with the particular category in which the job is placed), rather than the true aptitude profile unique to that job. Again, this is a situation in which a general impression of the job (in this case the characteristics of the category in which the job is placed) influences individual category ratings. Thus, several aspects of the conditions under which DOT ratings were obtained seem to correspond with those that often produce halo.

Note, however, that the extent to which analyst ratings reflect illusory halo cannot be directly assessed with the current data set. Generally, illusory halo is said to exist when observed intercorrelations among dimensions are higher than the true levels of intercorrelation (Murphy & Reynolds, 1988; Murphy, et al., 1993; Murphy & Jako, 1989; Pulakos, Schmitt, & Ostroff, 1986). However, there does not appear to be any reasonable way to estimate the true levels of intercorelation among occupational ability requirement dimensions that could be used to compare dimension intercorrelations obtained from

analyst ratings. Although intercorrelations among incumbent ability test scores could be used as an estimate of true levels of intercorrelation among dimensions of *individuals' abilities*, they do not necessarily provide a reasonable estimate of true levels of intercorrelation among dimensions of *occupational ability requirements*. Without such an estimate, a direct assessment of analyst illusory halo is not possible.

Finally, in addition to cognitive biases or inaccuracies and the specific conditions that may have produced halo errors, raters' implicit theories may also give rise to rating inaccuracies. That is, analysts may hold certain implicit beliefs about the nature of occupational aptitude requirement dimensions and relationships among these dimensions. These implicit theories may then influence the rating process such that analyst ratings reflect the analyst's implicit theory in addition to (or instead of) occupations' actual aptitude requirements. For example, analysts may hold the implicit belief that, in terms of occupational ability requirements, verbal ability and numerical ability are negatively correlated (i.e., as the verbal ability requirements of a job increase, the numerical ability requirements decrease). Analysts might then rate occupations' ability requirements in accordance with this belief, even though it may not be rooted in reality.

Although there appears to be little research demonstrating this phenomenon in the job analysis literature, the idea that implicit theories influence judgments and ratings has received support in areas such as performance appraisal (e.g., Cooper, 1981; Krystofiak, Cardy, & Newman, 1988) and personality (e.g., Passini & Norman, 1966). It seems likely that this general effect also holds for job analysis ratings. Job analysts may hold implicit beliefs about occupational requirements that influence their ratings of those requirements.

Analyst Profiles: Conclusion

From this discussion of the source and process involved in producing these ratings, it is difficult to specify precisely the nature of the job analyst data or how these data might be expected to behave when subjected to clustering. However, three broad conclusions seem reasonable. First, given all the potential sources of error outlined above, it is likely that these ratings contain substantive sources of error. Any (or all) of the six cognitive processes discussed can introduce or perpetuate error, providing ample opportunity for ratings to be less than accurate.

Second, many of the analyst profiles likely suffer from illusory halo to some degree. This situation appears likely, not only because of several of the conditions involved in obtaining DOT ratings, but also because of the categorization process involved in encoding and storing job-relevant information. However, even given halo, one still sees patterns of ability requirement ratings that make theoretical sense.

Third, many of the ratings may have been influenced by analysts' implicit theories about occupational aptitude requirements. To the extent that these theories were inaccurate, this influence may have introduced further error into analyst ratings.

Again, it is difficult to say what effect these factors might have on job clusters resulting from these data. Obviously, these errors and biases will affect the job clusters resulting from these data, but the nature of this effect depends upon the extent, consistency, and nature of the errors. Although it is possible to describe the nature of some of the biases that could be expected to be present (e.g., halo), it is more difficult to determine the extent and consistency of the bias. As a result, it is difficult to say whether these errors will have any substantive impact on resulting job clusters.

For instance, although halo error is likely to be present in analyst ratings to some extent, it is not clear how influential the halo effect might have been, or how consistent this effect was across raters and/or jobs. If the halo effect were consistent across raters and jobs, then this bias may not have much of an effect on job clusters (i.e., job clusters resulting from these data may not be substantially different from clusters resulting from data not subject to this bias). This is because a consistent halo bias might keep relative differences between jobs intact, although the sizes of the differences would be reduced. By keeping relative differences intact, data influenced by a consistent halo effect may still result in clusters similar to those that would result from data not influenced by this effect. On the other hand, if the halo effect was not consistent across raters and jobs (e.g., certain raters produced more halo errors, certain types of jobs were more subject to these errors), then this effect might have more of an influence on subsequent job clusters. However, the nature of this influence cannot be specified without knowing more about which occupations' ability profiles contain more or less halo error.

Similarly, although analysts' implicit theories might have influenced ratings, it is difficult to determine what effect this might have on job clusters without knowing the nature of these theories and how consistent they are across raters. For example, if most raters held similar theories, then we might predict that job clusters resulting from these data would reflect the nature of these theories. However, without knowing the nature of the theories, it is difficult to make a more specific prediction about the nature of resulting job clusters. On the other hand, if raters held idiosyncratic theories, these theories may have introduced unsystematic error into the data. Such a situation would also make it

difficult to produce any specific predictions about the influence of implicit theories on resulting job clusters.

Estimating Ability Requirements Through Incumbent Testing

Information Source

Another method of estimating jobs' ability requirements is to test incumbents performing satisfactorily. Obviously, for this method, incumbents are the reasonable and direct source of ability requirements data. Generally, the idea is that if a given worker is performing satisfactorily on the job, then that worker's levels of abilities must be sufficient for any worker's satisfactory performance. Thus, the ability test scores (i.e., ability levels) of this type of worker should be a good representation of the job's ability requirements.

Incumbent scores were obtained with paper-and-pencil subtests of the GATB. Generally, paper-and-pencil tests administered in a group setting are superior to individually-administered tests in terms of standardization and efficiency (Murphy & Davidshofer, 2001). Paper-and-pencil tests present the same stimulus to each person, are conducted under the same or similar testing conditions, are scored objectively, and can be administered to many individuals at the same time. On the other hand, these types of tests are inferior to individually administered tests and computerized adaptive tests in terms of their ability to collect data on test behavior and to tailor questions to subjects' previous response patterns (Murphy & Davidshofer, 2001). Other than the answer given, paper-and-pencil tests often do not provide information about test behavior that might be relevant in determining test-takers' ability levels, such as why an incorrect answer was chosen. In addition, these tests do not use information about test-takers to tailor

questions to subjects, another potential disadvantage when attempting to measure ability levels accurately. However, in general, well-developed paper-and-pencil ability tests tend to have good psychometric characteristics and provide reasonable estimates of test-taker ability levels.

The GATB is a well-researched ability battery. The long history of research on this aptitude battery indicates that it is a valid measure of respondents' ability levels (e.g., U.S. Department of Labor, 1970; 1979; Bemis, 1968; Droege, 1968; Knapp, Knapp, & Michael, 1977; Hakstian & Bennet, 1978). Thus, it seems reasonable to conclude that the incumbent ability profiles used in this study constitute reasonably valid measurements of incumbents' levels of abilities.

Process

The process of obtaining incumbent ability test scores involves the administration of one or more ability tests. Rather than involving the process of human cognitive estimation of required abilities as discussed previously, this process involves a more direct measurement of worker abilities. Through administering one or more aptitude tests, samples of behavior (e.g., verbal, spatial, numerical) are obtained from incumbents. These samples are then interpreted as indicators of incumbents' underlying abilities. Mean incumbent scores for a given job are then interpreted as indicators of reasonable occupational ability requirements.

Note that both test-score data and analyst data represent indirect estimates of jobs' ability requirements. However, the bases for these indirect estimates differ. Incumbent testing involves measuring current incumbent ability levels to estimate occupational ability requirements. Although in many cases, present employee ability levels are

reasonable estimations of job ability requirements (e.g., because present employees have been selected because they have the required level of a particular ability), the process of job incumbent testing nonetheless results in *indirect* estimates of occupational ability requirements based on *current average* employee ability levels.

Similarly, analyst ratings of a job's ability requirements are also indirect estimations of an occupation's ability requirements, but in this case these indirect estimates are more likely to be based on job analysis data consisting of observations of job tasks and how these tasks are completed, rather than measurements of current employees' ability levels directly. Thus, although both the rating and testing processes result in indirect estimates of ability requirements, the *bases* for these indirect estimates differ between the two processes (i.e., analyst data are task-driven whereas incumbent data are test-driven).

Constructs. Although incumbent testing and analyst rating represent different methods of measurement, for the present purposes the goal of measurement is the same: to determine a job's ability requirements (or the ability levels that workers bring to the job). In this sense, the intent is to measure the same constructs. Scores from both datasets are intended to measure general aptitude requirements for average, satisfactory performance.

However, because the methods used to measure these job requirements are quite distinct, each type of data will have its own unique practical and theoretical strengths and weaknesses as an indicator of these aptitude constructs. These unique aspects of each type of data may produce differing job clusters, which in turn may affect outcomes associated with the use of these clusters.

Bias/Inaccuracy. Errors in estimating aptitude requirements through incumbent testing can be introduced in several ways. Generally, these errors in measurement occur when factors other than incumbents' ability levels influence test scores. For instance, some research suggests that performance on cognitive ability tests is affected by test-taker motivation (Arvey, Strickland, Drauden, & Martin, 1990), even when the effects of previous test performance are controlled (Chan, Schmitt, DeShon, Clause, & Delbridge, 1997). Chan, et al. (1997) found that, through its influence on test-taking motivation, face validity perceptions of a cognitive ability test also affected performance. Other research indicates that cognitive ability test performance can be influenced by test-takers' awareness that, by taking the test, they risk confirming a negative stereotype about a group to which they belong. For example, African American test-takers may be at risk for confirming a negative stereotype about this group's cognitive ability, or female test-takers may be at risk for confirming a negative stereotype about this group's mathematical ability (Steele & Aronson, 1995; Steele, 1997). These are but a few of the many factors involving the test-taker and the test-taking situation (e.g., distractions) that may introduce error into test scores. It is difficult to say which of these factors were influential in the GATB test score database, but assuming that, generally, standardized guidelines for administering this test were followed, the influence of irrelevant factors may have been minimized.

Because incumbent test scores estimate general job-level aptitude requirements, error can be introduced through incumbent sampling. Any given sample of job incumbents may be unique or biased in some way, particularly if the sample is small and/or the job idiosyncratically represents the occupation. Mean scores from this type of

34

sample are biased estimates of occupational aptitude requirements. Thus, to the extent incumbents are not sampled appropriately, sampling error will also affect aptitude estimates.

Examining the process of incumbent testing as a measurement of the abilities of a sample of incumbents reveals another potentially important source of error in these scores. This source of error stems from two factors: (1) the likelihood that, for most jobs, only a few aptitudes from the full GATB profile are actually essential to performance on the job, and (2) the well-known tendency for ability dimensions to be positively correlated (Carroll, 1993; Spearman, 1904; U.S. Department of Labor, 1970).

It seems likely that for most jobs, only a subset of abilities from the full ability profile is *essential* for the job, where a particular level of these abilities is required for satisfactory performance. The rest of the abilities, then, are *nonessential* for the job, where incumbents could have a wide range of levels of these abilities and still perform satisfactorily. For example, for the job of computer programmer, numerical ability would probably be considered essential, whereas motor coordination would be probably considered nonessential.

Having established this distinction, it appears that, when using incumbent test scores, estimates of nonessential abilities are more likely to be inaccurate than estimates of essential abilities, in part, because of selection and turnover processes. Essential abilities are likely to be directly selected for in most organizations, and thus incumbents' ability levels on these types of abilities are likely to accurately reflect at least minimal job ability requirements. In addition, incumbents with less than the required levels of abilities essential for satisfactory performance are more likely to quit or be fired, and

those with more than required levels may also tend to "gravitate" toward other more appropriate and challenging employment (Wilk, Desmarais, & Sackett, 1995; Wilk & Sackett, 1996). In short, selection and turnover processes function to increase the probability that incumbent test scores on essential ability dimensions will accurately reflect occupations' ability requirements for these essential ability dimensions.

Clearly employers are less likely to hire or fire employees based on nonessential abilities. For example, it would be unwise for organizations to select (or fire) computer programmers based on their levels of motor coordination. Thus, selection and turnover processes do not function to increase the probability that incumbent test scores on nonessential ability dimensions will accurately reflect occupations' ability requirements.

The probability that incumbent scores for nonessential ability dimensions will be inaccurate reflections of actual job requirements is further increased by the second factor mentioned previously: the tendency for scores on ability tests to be positively correlated. This tendency implies that when incumbents are measured along multiple ability dimensions, they will tend to receive similar scores on both essential and nonessential. For example, if a high level of numerical ability is selected for when hiring programmers, the natural tendency for this ability to be nontrivially correlated with motor coordination (or other ability dimensions) will increase the likelihood that incumbent programmers' scores on tests of motor coordination will also be relatively high (indicating that the job requires relatively high levels of this ability), even though the job may not require this ability at all. Thus, some error may be introduced when required levels of nonessential abilities are estimated using mean incumbent ability test scores.

This tendency for ability dimensions to be positively correlated also implies that both analyst and incumbent data may tend to demonstrate a halo effect. That is, ability profiles from both sources may tend to reflect less within-job variability than actually exists for most occupations. The conditions under which DOT ratings were obtained, and the nature of the cognitive categorization process and implicit theories seem to point toward the possibility that the analyst profiles reflect the halo effect. Similarly, the natural tendency for ability test scores to be positively correlated also points toward the strong possibility that mean incumbent profiles reflect some sort of halo effect.

Incumbent Profiles: Conclusion

Three conclusions seem reasonable from this discussion. First, in this context, both analyst ratings and incumbent ability test scores are intended to measure the same constructs: aptitude requirements for average, satisfactory performance. However, because the methods used to gather these data are so dissimilar, each type of data likely has unique strengths and weaknesses that may affect resulting job clusters and some of the outcomes associated with the use of these clusters. Second, estimating ability requirements through incumbent testing may lead to relatively less accurate estimates of nonessential ability requirements (in comparison to estimates of essential abilities). Finally, although for different reasons, profiles resulting from both analyst rating and incumbent testing may often reflect a halo effect to some degree.

Estimating Ability Requirements Through Regression

Information Source

Another potential method of developing occupational ability requirement profiles involves statistical estimation. For this method, both ability requirement data and other

37

occupational data are used to develop prediction equations. These equations can generate predicted ability requirement scores from other occupational data when jobs lack ability requirement data.

This type of method was used in the development of O*NET's Ability Profiler. In this case, the 'other occupational data' used to predict GATB ability requirements consisted of other DOT job analysis data. DOT ratings of several diverse aspects of jobs (see Table 1), as well as incumbents' GATB scores, were used to develop prediction equations. These equations were then used to predict GATB aptitude requirements for those DOT jobs in which incumbents were not tested with the GATB. Thus, regression-estimated ability profiles used in this study are based on job analyst ratings. That is, ratings of occupational characteristics and requirements such as specific vocational preparation, temperaments, and physical demands constitute the source of these ability requirement estimates. Through a process of data reduction and statistical prediction, these data were used to develop ability requirement profiles for each DOT job.

Process

As discussed previously, the process involved in the development of regression-estimated ability profiles involved both principal component analysis and regression analysis. Principal component analysis was used to reduce the numerous DOT job analysis variables to a set of seven promax-rotated components. Mean incumbent GATB scores for each occupation were then regressed on these component scores. This yielded a set of regression weights that could be applied to the set of component scores for each job to estimate each DOT occupation's GATB ability requirements.

38

Thus, this process, like those involved in incumbent testing and analyst estimation, results in indirect estimates of occupational ability requirements. However, the manner in which these indirect estimates were developed differs. Specifically, regression-estimated profiles can be seen as a final product of both the analyst estimation and incumbent testing processes, as well as a statistical estimation process. Analyst estimation and incumbent testing produced the data subjected to regression analysis. However, although the regression-based data are in part the result of processes involved in analyst estimation and incumbent testing, the regression-estimation process as a whole differs substantially. In particular, although the regression analyses were based on analyst and incumbent data, these data were mechanically combined in a statistical estimation process in which regression weights were applied to these job characteristics data.

This difference may have resulted in noticeable differences in estimates of ability requirements for many jobs. For example, the regression estimation process does not allow factors other than the given set of occupational data to influence ability estimates. Analyst estimates, however, can be influenced by a variety of other factors (e.g., other occupational information, cognitive limitations). On the one hand, this may mean regression estimates tend to be more accurate; they are free from the cognitive biases and limitations present in analyst estimation. On the other hand, this may mean analyst estimates tend to be more accurate; analysts have the flexibility to incorporate other, potentially important, occupational information into the process and to weigh the information presented to them in unique ways, depending upon the circumstances. The bulk of research on the broader issue of clinical versus actuarial prediction indicates that

regression-estimated profiles tend to be superior (e.g., Dawes, Faust, & Meehl, 1989; Marchese, 1992; Sawyer, 1966; Meehl, 1954), though dissenting opinions exist (e.g., Holt, 1970). In addition, some research (e.g., Sawyer, 1966) suggests that an effective strategy may be to use skilled observers to collect information and then use actuarial methods to model and formulate consistent predictions or interpretations from this information (Dawes, et al., 1989). Given that a similar procedure was used to develop the regression-estimated ability profiles used in this study, we might expect these profiles to be better estimates of occupational ability requirements than those developed through analyst ratings.

Similarly, differences between the process used to develop incumbent score profiles and the process used to develop regression profiles will also likely result in substantial profile differences. Although incumbent ability data were used in the development of the regression weights used to estimate ability requirements, the general approaches involved in these methods are very different. Whereas the incumbent testing process uses information about individuals (i.e., incumbents' ability levels) to estimate ability requirements, the regression-estimation process uses information about occupations (i.e., DOT data) to estimate these requirements. As discussed in depth previously, these sources can lead to differing ability estimates. In addition, incumbent profiles were obtained from a testing process, whereas regression profiles were developed through a statistical estimation process. The fact that these processes are substantively different, and thus have unique advantages and limitations, also increases the likelihood that ability estimates differ between the two methods.

Constructs. Again, the goal of developing regression-estimated profiles was to estimate occupational ability requirements. Scores resulting from this estimation process, as well as those resulting from analyst rating and incumbent testing, are intended to measure general aptitude requirements for average performance. However, as mentioned before, because the methods used to measure these job requirements are dissimilar, each type of data may have its own unique strengths and weaknesses. Again, this is not to say that one type of data is 'best' overall, but rather that each type of measurement may have unique strengths and weaknesses, making it more or less useful for a given purpose.

Bias/Inaccuracy. On one hand, we might expect the sources of potential error in scores obtained from regression estimation to be quite different from the potential sources of error in scores obtained from analysts and incumbents. Because regression-estimated scores are obtained through a statistical estimation procedure, it should be not subject to the types of biases present in analyst and incumbent scores (e.g., halo effects). On the other hand, however, both analyst and incumbent data were used in the development of the regression equations, and analyst data were used as the basis for prediction. Thus, any error present in these data will influence scores obtained from the regression equations.

Specifically, using mean incumbent test scores as the criterion to develop prediction equations seems to indicate that regression-based scores will contain error in two ways. First, regression-based scores will contain prediction error because prediction is imperfect. Second, because mean incumbent test scores contain error due to particular sampling of jobs and incumbents, even if perfect prediction were possible, estimates would still contain some random error in terms of accurately reflecting occupational

ability requirements. In attempting to predict test scores, then, regression equations may tend to produce errors similar to those found in the test-score data perhaps most notably, something similar to halo error.

In addition, because analyst data from the DOT were used as predictors, any error contained in these data will also affect regression-estimated scores. If these data contain systematic errors, these errors may also show up in regression-estimated scores.

Regression Profiles: Conclusion

Two conclusions can be drawn from this discussion. First, prediction errors in regression-based scores may contain three components: errors in prediction, errors stemming from incumbent data, and errors stemming from analyst data. Given the likelihood that many analyst and incumbent profiles are subject to halo error, we might expect regression-based data to demonstrate this type of error as well. However, as was the case for the other types of data, without knowing the extent and consistency of these types of errors, it is difficult to say what effect this will have on job clusters.

Finally, the previous sections have asserted that each type of data is intended to measure the same thing: satisfactory levels of each cognitive ability. In detailing this assertion, I have also outlined the strengths and biases of each type of data. With all that said, it is important to re-emphasize that there are no 'bests': no 'best' data for describing satisfactory ability (see Campbell & Fiske, 1959, for the multi-trait multi-method approach to construct validity), no 'best' clustering algorithms, and no 'best' criteria by which to gauge the job-cluster solutions. On the other hand, there are many 'usefuls' where this study will make advances, both on conceptual and empirical levels: the most useful data for describing satisfactory ability, the most useful clustering algorithms, and

the most useful criteria by which to gauge the job-cluster solutions. The following sections discuss this further.

Implications for Personnel-Related Functions

This discussion of the sources and processes involved in the development of analyst, incumbent, and regression ability profiles indicates that substantial differences may exist in the ability estimates resulting from these different methods. It follows, then, that these three types of data may not result in the same (or even similar) job clusters, which has implications vis-à-vis the purposes to which the clusters are put, such as test validation or job evaluation. As mentioned previously, job clusters underlie many personnel-related activities, so any differences in job clusters will most likely produce differences in the outcome of these functions. So, creating profiles of job descriptors may require choosing both the broad psychological domain and quantitative method carefully according to the objective of classification. In addition, the choice of type of data within that domain may be important, so that the cluster analysis can yield useful job clusters.

Job clustering or classification serves a number of important purposes in organizations (see Table 3). However, job clusters based on ability requirements are only appropriate for some of these purposes. Specifically, ability-based job clusters might be appropriate when jobs are classified for test validation, vocational and educational guidance, job placement, personnel classification, internal job classification, job evaluation, and exploratory research, theory development, and methodological research objectives. The nature of each of these objectives seems to indicate that job classifications based on similarities and differences in ability requirements would be

appropriate. For example, ability-based job families might be useful in vocational guidance situations because job seekers can take ability tests and then focus their searches within clusters of jobs that match their ability test score profiles. In addition, it may be desirable to cluster jobs according to ability requirements in order to validate ability-based selection tests.

This study can evaluate clusters resulting from the three different types of ability data described previously by examining how differences in these clusters might have implications vis-à-vis the purposes to which ability-based clusters are put. That is, relative strengths and weaknesses of analyst-based, test-based, and regression-based ability data can be evaluated by examining the effectiveness of clustering solutions resulting from these data for these purposes. However, this is only possible for those purposes for which criteria are available. For the sample of jobs at hand, it appears that criteria for examining the effectiveness of job clusters are available for only two major purposes: test validation (personnel selection) and job evaluation. Therefore, we can evaluate the effectiveness of clusters resulting from test-based, ability-based, and regression-based ability data for these two purposes.

Test Validation

In many situations, ability-based job clusters might be useful for selection test validation (Arvey & Mossholder, 1977). For example, it may be necessary to combine several jobs with similar ability requirements into larger job families in order to have a large enough sample for validation. In addition, even in situations in which sample size is not a concern, combining jobs with similar ability requirements may still be desirable. Instead of developing and validating several distinct selection tests for superficially

different jobs, organizations can cluster jobs according to similarities in ability requirements and develop and validate a smaller number of tests for these job families, simplifying the development, validation, and implementation of ability-based selection tests.

In order to examine the relative strengths and weaknesses of clusters resulting from analyst-based, test-based, and regression-based data for use in test validation, these sets of clusters can be compared in terms of the criterion-related validity coefficients associated with the jobs in each cluster. Specifically, for test validation purposes, it would be desirable to have job clusters consisting of jobs with similar criterion-related validities. Clusters consisting of jobs that have widely varying criterion-related validity coefficients would be inappropriate for test validation situations because these clusters would mask important between job differences in predictor-criterion relationships. Use of these types of clusters might lead to incorrect conclusions in the validation process. For example, use of these clusters might cause researchers to conclude that a predictor is not valid for all the jobs in a cluster, when in fact it is valid for some jobs in that cluster. On the other hand, use of these clusters could also lead to the conclusion that a predictor is valid for all the jobs in a cluster, when in fact it is not valid for some jobs in that cluster. Thus, for test validation purposes, the most useful clusters are those that consist of jobs with similar predictor-criterion relationships (and therefore, similar criterion-related validity coefficients). Therefore, the utility of analyst-based, test-based, and regression-based clusters for test validation can be compared by examining the amount of variability in criterion-related validity coefficients across cluster solutions: more within-cluster variability indicates less utility.

Job Evaluation

Ability-based job clusters may also be useful for job evaluation. Job evaluation is "a systematic procedure designed to aid in establishing pay differentials among jobs" (Milkovich & Newman, 1990, p. 595). Many types of information might be useful for determining occupational pay levels; ability requirement data is one example (e.g., see Milkovich & Newman, 1990). For instance, it may be appropriate to determine salary, in part, based on ability requirement levels such that individuals in jobs requiring higher ability levels receive higher pay. Ability-based job clusters could be used in this situation to determine which jobs should be paid similarly. That is, jobs are clustered according to ability requirements and the jobs within each cluster are paid similarly because they require similar levels of abilities. Therefore, the most useful ability clusters for job evaluation would be those consisting of jobs with little within-cluster variability in pay rates. Clearly, clusters containing jobs with widely varying pay levels would be inappropriate in these situations.

Therefore, the usefulness of job clusters based on analyst-, test-, and regression-based data for job evaluation can be compared by examining variability in pay levels across cluster solutions. Data on occupational pay levels are available from the Bureau of Labor Statistics (Bureau of Labor Statistics, 2000). These data indicate average (and median) hourly, monthly, and yearly pay rates. Again, less within-cluster variability in pay levels indicates a more useful cluster solution for job evaluation purposes.

Career Exploration

Finally, differences in job clusters based on these three types of ability data may also have implications for career exploration, in particular with O*NET's Ability

Profiler. If clusters obtained from the actual GATB profiles are substantially different from those obtained from the regression-estimated profiles, this may indicate that the regression-based profiles used by the Ability Profiler are not functioning the same way actual GATB profiles would in terms of the types of jobs Ability Profiler clients are encouraged to pursue. That is, if the Ability Profiler used actual GATB profiles rather than the estimated profiles, the types of jobs retrieved by the Profiler for a particular client might not be the same as those it would currently retrieve for the same client. Although this would not necessarily mean that the O*NET Ability Profiler is generating inappropriate suggestions for clients, it is worthwhile to know whether the Profiler might function *differently* if it included a different type of ability data.

In summary, the present analysis reveals differences and similarities in job clusters using three different types of ability data: job analyst data, incumbent test score data, and regression-estimated data. Relative strengths and weaknesses of analyst-based, test-based, and regression-based ability data are evaluated by examining the effectiveness of clusters solutions resulting from these data for three purposes: test validation, job evaluation, and career exploration. Specifically, within versus between cluster variability in criterion-related validity coefficients and pay rates are examined to evaluate the three types of data for test validation and job evaluation purposes, respectively, with large between-cluster and little within-cluster variability indicating useful clusters. In addition, similarities and differences in clusters resulting from these three data types are examined to reveal potential implications for career exploration, particularly with O*NET's Ability Profiler. Other analyses would certainly be possible, as there are no tidy prescriptions for

47

profile matching, but past research seems to indicate that this study represents one

reasonable way to analyze these issues.

METHOD

Research on quantitative clustering methods indicates that no single technique is superior in all situations, although a few tend to perform well relatively consistently (e.g., see Harvey, 1986; Milligan & Cooper, 1987; Colihan & Burger, 1995; Milligan, 1981; Milligan & Cooper, 1985). In particular, past research recommended three decision rules for determining the number of clusters present in the data, the cubic clustering criterion (CCC; developed by Sarle, 1983), the variance ratio criterion or pseudo F statistic (developed by Calinski & Harabasz, 1974), and the pseudo t^2 statistic (based on a statistic developed by Duda & Hart, 1973), and one clustering method, Ward's (1963) minimum variance technique.

In addition, each of these decision rules and the clustering technique appear to be appropriate for job clustering carried out for both test validation and job evaluation purposes (the purposes for which job clusters developed in this study are evaluated). Although no particular clustering procedures appear to be uniquely suited to clustering for test validation and/or job evaluation, these techniques are certainly reasonable choices for these purposes. Thus, given that these techniques are at least as appropriate as any others for the present purposes, and that previous research indicates they are often among the top performers, using the CCC, pseudo F, pseudo t^2, and Ward's technique is a reasonable way of getting convergent evidence on job clusters for this study.

Methods for Determining the Number of Clusters Present

The CCC, pseudo F, and pseudo t^2 were chosen to determine the number of clusters because previous research has indicated that these three methods are superior to most others. Milligan and Cooper (1985) found via simulation studies that among a set

49

of 30 procedures for determining the number of clusters in a data set, the CCC, pseudo F, and a statistic that can be transformed into a pseudo t^2 were among the best performers in terms of accurately identifying the number of clusters in the data, as specified by the simulation parameters. Each of these indices incorporates several types of statistics (e.g., within cluster sum of squares) for each step in the hierarchical clustering process to produce an index of cluster solution adequacy. Generally speaking, the indices can be used to determine whether two clusters joined at a given step in the clustering process should in fact be combined. Examination of values across hierarchical clustering steps can then inform decisions regarding the number of clusters present in a given data set.

Based on this previous research, others (e.g., SAS Institute, 1999) have suggested looking for agreement among these three statistics to determine the number of clusters present. Specifically, these criteria should be examined for local peaks in the CCC and pseudo F statistic as well as a small value for the pseudo t^2 followed by a larger t^2 at the next clustering step. This type of pattern indicates that the appropriate number of clusters has likely been identified.

Clustering Methods

Ward's method forms clusters by minimizing the total within-group or within-cluster sums of squares (i.e., the sum of the squared deviations of the scores about their mean). That is, clusters are produced at each step so that the resulting cluster solution has the fewest within-cluster sums of squares. This clustering method was also chosen based on past research. Several studies (e.g., Milligan & Cooper, 1987; Colihan & Burger, 1995) have indicated that Ward's method is a superior choice under many circumstances.

Data

All three types of ability data were developed at the DOT level. However, as discussed previously, O*NET and all government agencies collecting occupational information either are or will soon be using the SOC system. These two systems differ primarily in that the SOC is a broad classification system containing approximately 820 occupational classifications that subsume the DOT, which contains over 12,000 occupational classifications. Given that the SOC system will be used by government agencies for all future occupational data collection, any information obtained about occupations at the DOT level may be seen as outdated. To deal with this problem, occupations were clustered at both the DOT and SOC level. This way, information consistent with the more current classification system could be obtained, but these results could also be compared with results obtained using the original ability profiles (not aggregated profiles).

Missing data reduced the working data set from 545 to 518 DOT-level occupations. In addition, aptitude G (General Intelligence) was excluded from analyses because it is redundant with GATB aptitudes V, N, and S. Therefore, for 518 DOT-level jobs, analyses were conducted on actual, regression-estimated, and analyst-estimated ability profiles consisting of the 8 remaining GATB aptitudes – V, N, S, P, Q, K, F, and M (see Table 1).

SOC-level analyses were conducted on DOT-aggregated ability profiles. Profiles were generated by first using a DOT-SOC crosswalk to place the 518 DOT-level jobs into their 264 corresponding SOC categories. DOT-level ability profiles were then averaged within each SOC classification, yielding actual, regression-estimated, and

analyst-estimated ability profiles at the SOC level, each type of profile consisting of the same 8 aptitudes. The percentage of DOTs with data out of all DOTs that fit in each SOC (according to the DOT-SOC crosswalk) ranges from .4% (1 DOT-level occupation with data out of 251 DOT occupations that fit into the SOC) to 100% (1 DOT-level occupation with data where only 1 DOT occupation fits into the SOC), with an average of 30% and a standard deviation of 29%. Thus, the extent to which DOTs with data are representative of all DOTs within each SOC varies considerably across SOCs. Each type of ability data was analyzed separately at both the DOT and SOC level.

Table 4 presents the number of SOCs in the present dataset in each of the SOC Major Groups. To facilitate classification, the SOC system divides occupations into 23 Major Groups, 96 Minor Groups, and 449 Broad Occupations (Bureau of Labor Statistics, 2001). Occupations with similar skills or work activities are grouped at each of these three levels. As shown in Table 4, the SOCs in the present dataset cover 22 of the 23 Major Groups. Thus, although some types of occupations are better represented than others, the present dataset appears to contain a reasonable variety of occupational types, at least according to the SOC system.

RESULTS

Descriptive Statistics and Reliability

Tables 5, 6, and 7 provide means, standard deviations, and intercorrelations for actual test score, analyst, and regression estimated data at the DOT level, respectively. Tables 8, 9, and 10 provide means, standard deviations, and intercorrelations for actual test score, analyst, and regression estimated data at the SOC level, respectively.

Studies indicate that the GATB aptitudes are measured reliably across numerous situations and populations. For example, studies conducted with a variety of samples from high school, college, and adult populations using test-retest intervals of one day to three years generally produced reliability coefficients in the range of 0.80 to 0.90 (U.S. Department of Labor, 1970). Thus, the actual GATB test score data used in this study represent averages of what should be highly reliable incumbent test scores.

Unfortunately, less is known about the reliability of ratings available from the DOT, including the GATB ratings used in this study. As Miller, et al. (1980) note, "no checks appear to have been made of the validity and reliability of the ratings during the course of fourth edition production" (p. 169). However, a few researchers have attempted to estimate the reliability of these ratings by developing ratings using methods that duplicated as closely as possible the procedures used to generate DOT ratings and examining the reliability of these ratings (e.g., Cain & Green, 1983; Miller, et al., 1980; Geyer, Hice, Hawk, Boese, & Brannon, 1989). These studies indicated that although some scale reliabilities (defined as average interrater correlations) were above 0.70 (e.g., Data, People), others were rather low (e.g., Things estimates were around 0.45). Thus, reliability appears to be variable across scales. Most relevant to this study, Geyer, et al.

(1989) found that alpha coefficients for GATB aptitude ratings were greater than 0.80 when four raters were used (see Table 11 for a summary of their findings). This suggests the aptitude ratings available from the DOT are likely to be quite reliable.

The regression-estimated data are simply composites of DOT ratings. Therefore, these scores should have somewhat better reliability than the individual ratings. Although the reliability of individual ratings varies across scales, most seem to have acceptable levels of reliability (e.g., see Cain & Green, 1983). Thus, the reliability of the composites (i.e., regression-estimated scores) should be high.

Clustering Results

Number of Clusters

As discussed, the CCC, pseudo F, and pseudo t^2 were jointly examined to determine the number of clusters present in each dataset. Unfortunately, however, no clear solution appeared for any of the datasets. Generally, a clear solution exists when, at a given point in the hierarchical clustering process (i.e., when a certain number of clusters have been formed), these three statistics demonstrate agreement such that there exists a local peak in the CCC and pseudo F statistic as well as a small value for the pseudo t^2 followed by a larger t^2 at the next clustering step. Such clear agreement did not appear to exist for any of the datasets. For example, Figures 2, 3, and 4 present the CCC, pseudo F, and pseudo t^2, respectively, for actual test score data (SOC level) for 1 to 14 clusters. Although the CCC appears to have a few local peaks in this range, it is difficult to determine which should be chosen. In addition, the pseudo F does not appear to have any locals peaks, but rather rises fairly steadily from 14 clusters to 1 cluster. Also, although the pseudo t^2 has a few low points followed by larger values, again it is not clear which

local minimum should be chosen. Finally, and most importantly, examining these three figures jointly does not reveal a clear point at which these three criteria are met simultaneously. Thus, it is not clear how many clusters underlie this dataset. Similar problems occurred for each of the other datasets, although some of the datasets resulted in clearer solutions than others.

Therefore, because no single clear solution presented itself for any of the datasets, multiple cluster solutions were chosen for each dataset and used in subsequent analyses. That is, rather than somewhat arbitrarily selecting a single solution based on weak evidence, multiple solutions that seemed to be reasonable based on the CCC, pseudo F, and pseudo t^2 values were chosen. Results of subsequent analyses based on these solutions can then be examined in terms of convergence or differences across the multiple solutions. Specifically, for each dataset, a single solution was chosen for each of three ranges: (1) 2-14 clusters, (2) 15-34 clusters, and (3) 35-54 clusters.

These three ranges were chosen based on two considerations. First, it was assumed that for most practical purposes, somewhere between approximately 2 and 50 clusters would be the most useful and appropriate, given that approximately 500 occupations (at the DOT level) or 250 occupations (at the SOC level) were clustered. Second, for several of the datasets, peaks appeared in both the 2-5 cluster range and 15-25 cluster range. Therefore, choosing clusters solutions above and below the 15 cluster point seemed appropriate. Based on this, the three ranges were chosen to reflect a low range (below 15 clusters), a middle range (a range of 20 solutions starting the 15 cluster point), and a high range (a range of 20 solutions starting at the 35 cluster point) that would cover the 2 to 50 cluster range. The most appropriate solution in each of these

three ranges was chosen based on examination of CCC, pseudo F, and pseudo t^2 plots. Although in some cases this resulted in cluster solutions with a larger number of clusters than might be practical for some purposes, analysis of these solutions can still be used help understand how the three types of data behave in clustering situations. Table 12 presents the number of clusters chosen for each dataset based on these considerations. Subsequent analyses were conducted for each dataset using each of these three solutions.

Similarity of Cluster Solutions Across Data Types

The similarity of these cluster solutions can be compared across data types to examine the extent to which cluster solutions resulting from these types of data tend to agree. Specifically, cluster solution similarity can be examined by computing Hubert and Arabie (1985) adjusted Rand statistics. The Rand (1971) indexes the amount of agreement between two cluster solutions, indicating the extent to which pairs of occupations placed in the same cluster for one solution are also placed in the same cluster for the other solution and occupations placed in different clusters for one solution are also placed in different clusters for the other solution. However, the Rand statistic tends to be inflated because some of the agreement indexed by this value is to due to chance. Thus, Hubert and Arabie (1985) proposed an adjustment to the Rand designed to correct this statistic for the presence of chance agreement. The lower bound of the adjusted Rand is usually negative (depending on how the data are partitioned), with values near zero indicating that agreement between two sets of classifications can be attributed to chance (Milligan & Cooper, 1986). The upper bound of the adjusted Rand is 1.00. Milligan and Cooper (1986) found that this adjusted statistic functioned well as an index of cluster solution similarity in several situations.

Tables 13 and 14 present Hubert and Arabie adjusted Rand statistics for each cluster range at the DOT and SOC level, respectively. These tables reveal a couple of patterns. First, the three types of data clearly produce substantially different cluster solutions. None of the adjusted Rand values are above 0.45 and only two are above 0.20, indicating minimal cluster solution agreement above chance levels. Second, it appears that analyst and regression-estimated data tend to produce more similar cluster solutions than do actual test score and analyst data or actual test score and regression-estimated data. For all three of the cluster ranges at the DOT level and two of the three ranges at the SOC level, the analyst and regression-estimated comparison produced noticeably higher adjusted Rand values than did the actual test score and analyst comparison and the actual test score and regression-estimated comparison. However, although the analyst and regression-estimated comparison tends to produce relatively higher values, these values still appear to be quite small – the highest of the six is 0.42 and the other five are within 0.10 to 0.20. This indicates there is still substantial disagreement between analyst and regression-estimated cluster solutions.

Criterion-Related Validity Results

DOT Level Analyses

Criterion-related validity coefficients for the nine GATB aptitudes are available for each of the 518 DOT occupations included in this study. These values represent the correlation between GATB test scores and job or training performance. In most cases, the criteria were supervisory ratings, training ratings, or course/exam grades. Table 15 presents sample size means and standard deviations for the criterion-related validity studies. These values indicate studies examining G, V, N, and S tended to involve larger

57

samples than those examining P, Q, K, F, and M. Table 16 presents descriptive statistics for criterion-related validity coefficients at the DOT level.

The purpose of analyzing criterion-related validity coefficients was to examine the advantages and disadvantages of using the three types of data in job clustering for test validation purposes. Specifically, the extent to which occupational profiles of criterion-related validity coefficients are similar within clusters and differ between clusters for a given cluster solution may indicate the usefulness of that solution for test validation. For example, substantial within-cluster dissimilarity in occupational criterion-related validity coefficient profiles for a particular cluster solution could be seen as an indication that the cluster solution should not be used for test validation purposes. Using clusters consisting of jobs that have widely varying criterion-related validity coefficients would be inappropriate because these clusters would mask important between job differences in predictor-criterion relationships. Therefore, for test validation purposes, the most useful clusters are those that consist of jobs with similar predictor-criterion relationships (and therefore, similar criterion-related validity coefficients).

One way to examine variability in criterion-related validity coefficients across clusters is through profile analysis. Profile analysis is essentially an application of multivariate analysis of variance (MANOVA) to situations where several dependent variables (DVs) are measured on the same scale (Tabachnick & Fidell, 2001). These DVs can either represent several measurements of the same variable (i.e., repeated measures) or measurements of several different variables. In this case, the DVs are criterion-related validity coefficients for each of the nine GATB scales and therefore represent measurements of different variables.

58

Profile analysis addresses three major questions. First, this technique assesses the extent to which different groups (in this case clusters) have different mean DV levels. This "levels" test essentially examines the extent to which group profiles are similar in overall level. The null hypothesis for this test is that the mean of the means of the separate DVs is identical for the groups (Harris, 1975). Rejection of this hypothesis indicates that groups differ in terms of mean DV levels. In this case, the test involves examining the main effect for cluster membership. Mean criterion-related validities are averaged for each cluster. This average criterion-related validity variable is then tested for differences across clusters. Results indicate whether clusters are significantly different in terms of their average criterion-related validities, or in other words, whether clusters differ in terms of the level of their DV profile.

Second, this technique assesses whether the pooled profile for all the groups combined is flat (Harris, 1975). This "flatness" test essentially examines the extent to which scores on all DVs are similar. The null hypothesis for this test is that the average profile for all groups is perfectly flat (Harris, 1975). This hypothesis is rejected when, on average, scores for one or more DVs are higher (or lower) than scores for other DVs. In this case, the test involves examining the main effect for GATB scales. Criterion-related validities are averaged for each GATB variable. These average criterion-related validities are then tested. Results indicate whether certain GATB scales tend to be more or less predictive of criteria than other scales. Thus, whereas the levels analysis tests for between-cluster differences using means across aptitudes, the flatness analysis tests for between-aptitude differences using means across clusters.

Finally, this technique also assesses the extent to which different groups have parallel profiles of DVs. This "parallelism" test essentially examines the extent to which group profile shapes are similar. The null hypothesis for this test is that the profiles for the groups are parallel, meaning they have exactly the same shape (Harris, 1975). If this null hypothesis is rejected we conclude that the clusters differ significantly in terms of the shape of their criterion-related validity coefficient profiles. Here, the test involves examining the interaction between cluster membership and GATB scale. A significant interaction indicates the clusters' criterion-related validity profiles are not parallel (i.e., they are dissimilar in shape).

These three tests are therefore analogous to two-way analysis of variance tests (Harris, 1975). The levels test corresponds to a test of the cluster or group main effect, the flatness test corresponds to a test of the GATB aptitude main effect, and the parallelism test corresponds to a test of the interaction between cluster and GATB aptitude. Thus, the parallelism test generally takes precedence and qualifies any levels or flatness results.

Note that the following results are based on analyses that include criterion-related validity coefficients for G, which is a composite of V, N, and S. Analyses excluding G were also conducted. The pattern of results from these two sets of analyses was very similar. The largest difference between the two sets involved the flatness analyses. Results indicated that non-flatness of the validity profile accounted for more of the variance in validity coefficients when G was included than when it was excluded. For example, partial eta-squared values obtained for DOT level data indicated that between aptitude differences accounted for 49% to 58% of the variance when G was included, but

only 15% to 23% of the variance when G was excluded. However, the pattern of results remained the same and no other meaningful differences were observed across the two sets of analyses.

Tables 17, 18, and 19 present results of the "levels" test for cluster solutions in the 2-14, 15-34, and 35-54 cluster ranges, respectively, at the DOT level. It appears that actual test score data consistently produce clusters that differ significantly in terms of the level of their profiles of criterion-related validity coefficients ($p < .05$ in all cases). On the other hand, analyst data produce clusters that do not differ significantly in terms of validity profile level ($p > .4$ in all cases). Finally, regression-estimated data clusters had significantly different validity profile levels in the 15-34 cluster range ($p < .05$), but not in the 2-14 or 35-54 ranges ($p > .25$). Thus, actual test score data consistently produce significantly different mean validity profile levels whereas analyst data do not. Regression-estimated results are somewhat mixed, indicating these data tend to produce relatively similar validity profile levels, except in the 15-34 cluster range. This suggests that a middle range cluster solution was a better fit for the regression-estimated data in terms of criterion-related validity coefficient profiles.

Partial eta squared values (indicating the proportion of variance in averaged validity coefficients accounted for by cluster membership) also demonstrate this pattern. At all three cluster ranges, actual test score data produce clusters for which level differences account for more variance than do regression-estimated data; regression-estimated data, in turn, produce clusters for which level differences account for more variance than do analyst data. However, overall partial eta squared values appear to be relatively small. For example, cluster membership resulting from actual test score data at

the 2-14 cluster range accounts for only approximately 2% of the variance in averaged validity coefficients. In some cases values as high as approximately 20% (for actual test score data at the 35-54 cluster range) were obtained. In these cases it appears cluster membership is important, but nontrivial within-cluster variability in average validity coefficients remains and should not be discounted.

Tables 20, 21, and 22 present results of the "flatness" test for cluster solutions in the 2-14, 15-34, and 35-54 cluster ranges, respectively, at the DOT level. These results clearly show that the mean validity profile (across all clusters) is significantly different from flat ($p < .001$). This indicates that one or more of the GATB scales tend to be more (or less) predictive of criteria than the other scales. Partial eta squared values indicate non-flatness of the validity profile accounts for 49% to 58% of the variance. Thus, overall a substantial amount of variance is accounted for by this effect.

Figure 5 displays the mean validity profile across all clusters obtained from actual test score data analyses for the 2-14 cluster range solution at the DOT level. This mean profile is generally representative of those obtained from all other analyses (at both the DOT and SOC level) and indicates that, on average across jobs in the dataset, the General Intelligence (G) and Numerical Ability (N) tests tend to be the most predictive, whereas Verbal Ability (V), Spatial Ability (S), Motor Coordination (K), Finger Dexterity (F), and Manual Dexterity (M) are the least predictive, with Form Perception (P) and Clerical Ability (Q) falling somewhere in between. Results from post hoc analyses for this dataset (displayed in Table 23) confirm these observations. Specifically, these analyses indicate G, V, N, S, K, and F are significantly ($p < .05$) different from the mean validity

value, whereas P and Q are not. Overall, these results demonstrate the relative superiority of G and N as predictors of performance for the jobs used in this study.

Tables 24, 25, and 26 present results of the "parallelism" test for cluster solutions in the 2-14, 15-34, and 35-54 cluster ranges, respectively, at the DOT level. These results indicate there is a significant GATB scale by cluster membership interaction for all three types of data at all three cluster solution ranges ($p < .01$). These significant interactions indicate that in all cases mean validity profiles are not parallel. In other words, each type of data produces clusters that differ in terms of validity profile shape. Although partial eta squared values (indicating the proportion of variance accounted for by differences in profile shape) are fairly similar for results for each of the three types of data, across cluster solutions regression-estimated data produce the largest values, followed by analyst data, and then actual test score data. Together, these findings indicate that all three types of data result in significant interactions, but this interaction tends to account for more of the variance in criterion-related validity in regression-estimated clusters than in analyst or actual test score clusters. Overall, partial eta squared values ranged from relatively low (e.g., .05 for actual test score data at the 2-14 cluster range) to moderate (e.g., .26 for regression-estimated data at the 35-54 cluster range), indicating there is a considerable amount of within-cluster variance in profile shape.

It should be noted that these parallelism findings render the levels and flatness results less relevant. That is, the significant interaction indicates that although overall profile levels may differ across clusters, the magnitude and/or direction of the differences depends on which GATB scale is examined. In addition, the interaction indicates that at

least one cluster's validity profile is not flat, suggesting the flatness test is somewhat irrelevant.

SOC Level Analyses

Although the criterion-related validity coefficients were developed at the DOT level, the coefficients can be aggregated to the SOC level. Specifically, DOT-level profiles of coefficients were averaged within each SOC classification. SOC level analyses were then conducted on these aggregated coefficients. As noted previously, the SOC system will be used by government agencies for all future occupational data collection, and therefore any information obtained about occupations at the DOT level could be seen as outdated. Thus, SOC level analyses were conducted to provide information consistent with the more current classification system. SOC level results can also be compared with results at the DOT level to examine the consistency of findings. Table 27 presents descriptive statistics for SOC level coefficients.

Tables 28, 29, and 30 present results of the "levels" test for cluster solutions in the 2-14, 15-34, and 35-54 cluster ranges, respectively, at the SOC level. These findings are very similar to those obtained at the DOT level. Actual test score data produced clusters that differ significantly in terms of the level of their profiles of criterion-related validity coefficients at all three cluster ranges ($p < .01$ in all cases). Analyst data did not produce significantly different validity profile levels ($p > .25$ in all cases). Regression-estimated results approached significance at the 15-34 cluster range ($p = .068$), but not at the 2-14 or 35-54 ranges ($p > .14$). Thus, these results closely parallel those obtained at the DOT level: test score data consistently produce significantly different mean validity profile levels, regression-estimated data tend to produce similar validity profile levels, and

analyst data consistently produce similar validity profile levels. Again, this pattern can also be observed for the partial eta squared values. Overall, these values varied from relatively low (e.g., 0.04 for actual test score data at the 2-14 cluster range) to moderate (e.g., .32 for actual test score data at the 35-54 cluster range), again indicating nontrivial amounts of within cluster variance.

Tables 31, 32, and 33 present results of the "flatness" test for cluster solutions in the 2-14, 15-34, and 35-54 cluster ranges, respectively, at the SOC level. These findings also match those obtained at the DOT level. Again, the average GATB profile (across clusters) is significantly different from flat ($p < .001$ in all cases). One or more of the GATB scales appear to be more (or less) predictive of criteria than the other scales. Partial eta squared values vary from 0.585 to 0.653, suggesting this effect accounts for a fairly substantial amount of variance. As mentioned in the DOT-level analyses, mean validities indicated G and N tend to be the best predictor of performance for the occupations included in this study.

Tables 34, 35, and 36 present results of the "parallelism" test for cluster solutions in the 2-14, 15-34, and 35-54 cluster ranges, respectively, at the SOC level. These findings are similar to those obtained at the DOT level in that there is a significant GATB scale by cluster membership interaction for all three types of data at all three cluster solution ranges ($p < .01$). However, the results differ in terms of partial eta squared values. Whereas DOT level analyses demonstrated a consistent pattern in which regression-estimated data produced the largest values, followed by analyst data, and then actual test score data, SOC level analyses failed to reveal any consistent pattern. Again, partial eta squared values were fairly similar across data types, but in this case the rank

order of data types in terms of these values was not consistent across cluster solution ranges. Together, these results indicate that all three data types resulted in significant interactions, and these data types could not be distinguished in terms of partial eta squared values. As was found at the DOT level, partial eta squared values ranged from relatively low (e.g., 0.07 for actual test score data at the 2-14 cluster range) to moderate (e.g., 0.32 for regression-estimated data at the 35-54 cluster range), indicating the presence of a reasonable amount of within-cluster variance in profile shape.

Again, note that these findings make the levels and flatness results less relevant. As mentioned, the significant interactions indicate inferences regarding profile levels and flatness must be qualified.

Pay Data Results

DOT Level Analyses

Pay data were not available for DOT level jobs.

SOC Level Analyses

Pay data are available for 260 of the 264 SOC occupations included in this study. The data used in this study, available from the Bureau of Labor Statistics (BLS; 2000), represent median annual income for the year 2000. The BLS collects these data through the Occupational Employment Statistics (OES) program, which involves a yearly mail survey designed to estimate employment and wages for various occupations (Bureau of Labor Statistics, 2001). To collect wage data, employers are asked to report the number of employees in an occupation that earn each of a given set of wage ranges (e.g., how many individuals in a given occupation earn $35,360 to $44,719 per year). Generally, the OES program surveys approximately 400,000 establishments per year, and bases

estimates on three years of data, or 1.2 million establishments. However, the 2000 data are based on only two years due to the transition to the SOC system. Therefore, the wage data used in this study are based on a survey of approximately 800,000 establishments over the course of two years. Table 37 presents descriptive statistics for these data.

The purpose of analyzing pay data was to examine the advantages and disadvantages of using the three types of data in job clustering for job evaluation purposes. As mentioned, job evaluation refers to the process through which jobs or positions are ordered or assessed with respect to their value or worth to an organization. This process is often used in determining pay rates. Within- and between-cluster variability in pay rates was examined for each data type to reveal the extent to which job clusters resulting from each type would be useful for job evaluation. Relatively less within-cluster variability and more between-cluster variability were interpreted as indicating the presence of a relatively more useful cluster solution for job evaluation purposes.

Descriptives

Tables 38, 39, and 40 present pay rate descriptive statistics for actual test score data for 3 cluster, 26 cluster, and 39 cluster solutions, respectively. Tables 41, 42, and 43 present pay rate descriptive statistics for analyst data for 3 cluster, 21 cluster, and 40 cluster solutions, respectively. Tables 44, 45, and 46 present pay rate descriptive statistics for regression-estimated data for 4 cluster, 22 cluster, and 42 cluster solutions, respectively.

Boxplots

Figures 6, 7, and 8 present pay data boxplots for 3, 26, and 39 clusters,

respectively, resulting from actual test score data. Each of these figures seem to indicate

there is a reasonable amount of between-cluster variability in pay rates, but also some

degree of overlap across clusters. Results for the three cluster solution are most

interpretable, indicating pay rates increase from cluster 1 to cluster 3, although pay rates

clearly overlap across clusters. The solutions for 26 and 39 clusters are less interpretable;

no clear pattern emerges and most of the clusters overlap substantially.

Figures 9, 10, and 11 present pay data boxplots for 3, 21, and 40 clusters,

respectively, resulting from analyst data. These figures also demonstrate some between-

cluster differences in pay rates, but substantial cluster overlap. In this case, the three

cluster solution indicates decreasing pay rates from cluster 1 to cluster 3, but again

nontrivial overlap among these clusters appears to exist. Solutions for 21 and 40 clusters

demonstrate no clear pattern and contain substantial overlap, making interpretation

difficult.

Figures 12, 13, and 14 present pay data box plots for 4, 22, and 42 clusters,

respectively, resulting from regression estimated data. A very similar pattern occurs in

these plots. The four cluster solution indicates decreasing pay rates from 1 to 4 clusters,

with overlap across the clusters (particularly between the second and third cluster).

Again, substantial overlap in the 22 and 42 cluster solutions makes these solutions

difficult to interpret.

Overall, these boxplots do not seem to indicate that one type of data or one cluster

solution is more useful for job evaluation purposes. Although the three and four cluster

68

solutions led to more interpretable results, there appeared to be nontrivial overlap across clusters, indicating these are not likely to be the most appropriate cluster solutions for use in job evaluation. Among the three and four cluster solutions, regression estimated data perhaps led to clusters with slightly less overlap in pay rates than actual test score or analyst data (particularly if clusters 2 and 3 were combined). However, the difference does not appear to be large enough to advocate use of regression-estimated data over actual test score or analyst data for job evaluation purposes. Thus, no strong conclusions can be drawn from examination of these boxplots.

Intraclass Correlation Coefficients

A more direct way to examine pay data is to calculate intraclass correlation coefficients (ICCs). Although there are numerous versions of the ICC, essentially these coefficients give the ratio of the variance of interest (often between group variance) over the sum of the variance of interest plus error variance (often within group variance; Shrout & Fleiss, 1979). Thus, these coefficients estimate the proportion of total variance that is due to the effect of interest. In this case, ICCs can be used to examine the relative amounts of between to within cluster variance in pay rates. As discussed, ideally, job clusters developed for use in job evaluation would include little within cluster variance in pay rates relative to the variance between clusters. Therefore, relatively larger ICC values (indicating a large between cluster variance to total variance ratio) for a given cluster solution could be taken as an indication that this solution is relatively more useful for job evaluation.

Table 47 presents ICC values and 95% confidence intervals for actual test score data, analyst data, and regression-estimated data at all three cluster solution ranges

(confidence intervals were computed using formulas from Donner & Wells, 1986). These results indicate the three data types perform fairly similarly. Although examination of the ICC values suggests that actual test score and regression-estimated data tend to perform slightly better than analyst data, the values obtained are fairly similar and the confidence intervals show substantial overlap across data types. Therefore, there do not appear to be significant differences in performance across data types.

However, there do appear to be some differences across cluster solution ranges. Although the confidence intervals across cluster ranges overlap, those at the 2-14 cluster range tend to be much larger and include zero (or approximately zero), whereas those at the 15-34 and 35-54 ranges are smaller and do not include values below 0.25. These results suggest cluster solutions consisting of a larger number of clusters, where there is more opportunity for between-cluster variability and less opportunity for within-cluster variability may be more useful for job evaluation purposes.

Finally, it appears that overall, cluster membership accounts for moderate amounts of variance in pay rates. Specifically, ICCs ranged from 0.286 to 0.657. This suggests that although cluster membership does account for a reasonable amount of variance, there are nontrivial amounts of within cluster variability in pay rates. This variability may be of concern if clusters such as these are used in job evaluation situations.

DISCUSSION

This study examined similarities and differences in job clusters resulting from different types of occupational data. Previous research has indicated that different types of job characteristic data can produce substantially different job clusters (e.g., Ghiselli, 1966; Pearlman, 1980; Cornelius, Carron, & Collins, 1979). However, this research has tended to focus on occupational data from different psychological domains, comparing, for example, job clusters produced by task data with job clusters produced by ability data. This study took a different approach by comparing occupational data that belong to the same psychological domain (abilities), but differ in terms of the manner in which they were developed (e.g., analyst rating versus incumbent testing). In addition, the present study examined not only the extent to which job clusters resulting from these different types of data differ, but also the implications of such differences for some of the purposes to which job clusters are put. Specifically, relative strengths and weaknesses of three types of occupational ability data were evaluated by examining the effectiveness of clustering solutions resulting from these data for three purposes: test validation, job evaluation, and career exploration.

Data Types

As mentioned, three types of occupational ability data were examined in the present study. Specifically, job clusters resulting from actual test score, analyst, and regression-estimated ability data were compared. These three types of data represent estimates of the level of nine abilities (the GATB ability dimensions) required for performance in occupations. That is, each data type consists of estimates of occupational ability requirements along the nine GATB dimensions. Therefore, the three types of data

measure the same constructs: level of abilities required for average, satisfactory performance.

However, these data types differ in terms of the manner in which they were developed. Actual test score data were collected from satisfactorily performing incumbents; analyst data were obtained from a job analyst rating process; and regression-estimated data were developed through a process of statistical estimation. Thus, both the source of information from which each type of data was generated and the developmental processes involved differ across the three types. For example, actual test score data stem from satisfactorily performing incumbents and are obtained through an ability testing process, whereas analyst data stem from job analysts and are the result of a cognitive estimation process. These differences in development may then have resulted in substantially different estimates of ability requirements across data types, which in turn, may produce differing job clusters. These differences in job clusters across the three types of data may then have important implications for any personnel-related functions or activities using these clusters.

Implications for Test Validation, Job Evaluation, and Career Exploration

As discussed previously, job clustering underlies numerous personnel-related functions (see Table 3). Because it often plays an important role in these activities, clustering can influence their effectiveness. In addition, any differences in job clusters will most likely produce differences in the outcomes of these functions. Therefore, differences in job clusters resulting from the use of different types of data may have implications for the purposes to which clusters are put. This study examined these

potential implications for three purposes: test validation, job evaluation, and career exploration using O*NET's Ability Profiler.

Ability-based job clusters might be useful for selection test validation for a number of reasons (Arvey & Mossholder, 1977). For example, combining jobs with similar ability requirements may be necessary to obtain a large enough sample for validation. In addition, even when sample size is not a concern, organizations can cluster jobs to simplify the development, validation, and implementation of ability-based selection tests. In these situations, an ideal job cluster would be one containing occupations with very similar ability test score-performance relationships. A cluster containing jobs that have widely varying ability test score-performance relationships would be inappropriate for test validation situations because this type of cluster would mask important between job differences in predictor-criterion relationships. Therefore, different cluster solutions can be compared by examining the extent to which they exhibit within-cluster similarity in predictor-criterion relationship and between-cluster differences in these relationships. Relatively less within- to between-cluster variability could be seen as an indicator of a relatively more useful cluster solution.

Ability-based job clusters might also be used in job evaluation, the process through which occupational pay levels are determined. For example, ability-based job clusters could be used to determine which jobs should be paid similarly, where jobs are clustered according to ability requirements and the jobs within each cluster are paid similarly because they require similar levels of abilities. In this situation, an ideal job cluster would be one containing occupations with very similar pay levels. A cluster containing jobs that have widely varying pay levels would be inappropriate for job

73

evaluation situations because this type of cluster would mask important between job differences in pay rates. Therefore, different cluster solutions to be used for job evaluation purposes can be compared by examining the extent to which they exhibit within-cluster similarity in pay rates and between-cluster differences in these rates. Again, relatively less within- to between-cluster variability could be seen as an indicator of a more useful cluster solution.

Finally, differences in job clusters based on the three types of ability data may also have implications for career exploration involving O*NET's Ability Profiler. Specifically, if clusters obtained from the actual test score GATB profiles are substantially different from those obtained from the regression-estimated profiles, this may indicate the Ability Profiler - which uses regression-estimated profiles - might function differently if it used actual test score profiles (those profiles regression-estimated scores are intended to predict). Although this would not necessarily mean that the O*NET Ability Profiler is generating inappropriate suggestions for job seekers, it is worthwhile to know whether the Profiler might function differently if it included a different type of ability data.

<center>Findings</center>

<center>*Aptitude Intercorrelations*</center>

Before going into the specifics of the clustering results and their implications for test validation, job evaluation, and career exploration, a general characteristic of the GATB data should be discussed, namely the level of aptitude intercorrelation (shown in Tables 5-10). The eight GATB dimensions used in this study showed high levels of intercorrelation at both the DOT and SOC level, particularly for regression-estimated and

<center>74</center>

actual test score data. For example, actual test score data demonstrated an average aptitude intercorrelation of .63 (with a range of .32 to .90) at the DOT level and .65 (with a range of .34 to .89) at the SOC level. Regression-estimated data demonstrated an average intercorrelation of .85 (with a range of .64 to .98) at the DOT level and .80 (with a range of .55 to .97) at the SOC level. On the other hand, GATB scores obtained from analysts tended to be noticeably less correlated, demonstrating an average intercorrelation of .28 (with a range of -.24 to .71) at the DOT level and .24 (with a range of -.35 to .72) at the SOC level.

In addition, these average dimension intercorrelations include correlations between cognitive/perceptual aptitudes (e.g., verbal ability) and motor/dexterity aptitudes (e.g., manual dexterity). Although these two types of aptitudes are positively correlated, these relationships tend to be relatively low, as motor and cognitive abilities, though nontrivially related, are somewhat distinct. These relatively lower correlations between aptitudes from these two ability domains (motor and cognitive) then mask to some extent the high levels of intercorrelation within each of the two domains. In other words, although the average aptitude correlations described above (which include both motor and cognitive aptitudes) are quite high, the relatively low correlations between aptitudes from these somewhat distinct domains hide the even higher correlations that exist within each domain. For instance, by removing finger dexterity, manual dexterity, and motor coordination from consideration (the three aptitudes that are more motor in nature) average actual test score intercorrelation increased from .63 to .78 at the DOT level and from .65 to .77 at the SOC level; average analyst rating intercorrelation increased from .28 to .44 at the DOT level and from .24 to .34 at the SOC level; and average regression-

estimated score intercorrelation increased from .85 to .92 at the DOT level and from .80 to .89 at the SOC level. This suggests the cognitive/perceptual aptitudes included in the GATB are highly related, particularly when measured through testing and regression-estimation.

Overall, it is clear that the GATB aptitudes are highly related whether we focus on only the cognitive/perceptual dimensions or all eight aptitudes included in this study. In fact, several of the correlations appear to be at the limit of what the scale reliabilities will allow: the correlations are less than 1.0 perhaps only because the scale reliabilities are less than 1.0. This high level of correlation reflects a general cognitive ability factor (g) measured by all tests that require cognitive effort. It appears that the eight GATB dimensions, rather than measuring distinct attributes, are primarily measuring the same attribute (particularly when measured by test scores and regression-estimated scores): general cognitive ability.

This high level of aptitude intercorrelation is important because it places constraints on the clustering results that can be obtained from these data. In particular, it likely restricts the underlying cluster structure in terms of both the number of clusters that can exist and the manner in which occupations can be grouped. If the different GATB aptitudes are to some extent measuring only a single attribute (general cognitive ability) then occupations can only be differentiated or grouped according to this single ability factor. To the extent that the eight GATB aptitudes are measuring the same thing, each occupation's ability profile is effectively reduced to a single number reflecting the occupation's general cognitive ability requirement. Occupations can then be differentiated or grouped according to the level of their general ability requirement but

not the pattern (or shape) of requirements across different aptitudes. Therefore, occupational groups would differ only in that some groups require higher general cognitive ability levels than others. In this way, the manner in which occupations can be differentiated or grouped is constrained, which also likely restricts the number of clusters present.

Note that there is some differentiation among the aptitudes in the data used in this study, particularly the analyst data. However, the intercorrelations for both actual test score data and regression-estimated data are high enough that these data likely allow for differentiation among occupations primarily according to differences in general cognitive ability requirements, rather than differences in patterns of aptitudes. Recognition of this is important in understanding the clustering results and their implications. For instance, this indicates that for the most part any occupational clusters based on actual test score or regression-estimated data produced in this study should be interpreted as primarily the result of differences in general cognitive ability level rather than differences in patterns of any kind. This type of interpretation may then qualify some of the findings. For example, results pertaining to the 35-54 cluster ranges may be suspect as it seems unlikely that approximately 500 occupations (or 250 at the SOC level) can be grouped into 35-54 meaningfully different groups according only to differences in general cognitive ability requirements. This number of occupational groups might be more plausible if groupings were based on both shape and level differences (e.g., at one level there could be multiple shapes), but it seems unlikely that making this many distinctions based only on level differences is appropriate in most circumstances.

Aside from their implications for the results of this study, these findings regarding the intercorrelations among aptitude dimensions may also have broader implications for the differentiation among and/or grouping of occupations in general. Specifically, the size of the intercorrelations obtained suggests that when using mean test score data (or data designed to estimate mean test scores) from multiple aptitude batteries such as the GATB to describe occupational ability requirements, it may only be possible to meaningfully differentiate or group occupations according to levels of general cognitive ability (rather than patterns of ability requirements). The high intercorrelations among mean tests scores suggest the GATB measures little more than g; thus, when differentiating or grouping occupations using GATB profiles one has little more to work with than g.

This means that to some extent interpretations of GATB profiles as patterns of scores (i.e., more than just g) and efforts based on this type of interpretation (e.g., matching individuals to occupations according to ability patterns/shape) may be inappropriate. For example, the O*NET Ability Profiler (described previously) is designed to match individuals to occupations according to the similarity between the shape of the individual's GATB profile and the shape of occupational GATB requirement profiles (operationalized as the correlation between the individual's GATB scores and occupations' regression-estimated GATB scores). However, the extremely high correlations among the regression-estimated GATB scores indicate that scores on all eight dimensions are little more than repeated measures of g and thus occupational GATB profiles differ from flat mostly as a result of measurement error. Thus, using these scores to make conclusions about the similarity of shapes seems inappropriate.

However, these scores could be used to indicate the extent to which the individual and occupation are similar in terms of level of *g*. Similarly, GATB profiles could be used to group occupations according to differences in ability levels, but not differences in shape/pattern.

However, this does not appear to be true when analyst data are used to describe occupational ability requirements. Though substantial, the intercorrelations among GATB dimensions for analyst data were much lower, suggesting analyst scores do reflect different attributes. In other words, analyst scores appear to measure more than just *g*. It could be argued that the low correlations are simply a function of low reliability. However, although evidence regarding reliability of analyst data from the DOT is scarce, the available evidence suggests these data (particularly the GATB ratings) likely have an acceptable level of reliability (e.g., Geyer, et al. 1989). Thus, it appears that in contrast to test score data, analyst data do reflect distinct aptitudes to some extent. It may be that although the GATB subtests actually measure *g* primarily, analysts are able to distinguish among the dimensions to some extent. In any case, the fact that analyst profiles appear to measure more than *g* suggests these profiles could be used to differentiate or group occupations according to differences in both ability requirement level (*g*) and pattern/shape. Thus, for example, matching individuals to occupations or developing occupational groups according to similarity in GATB profile patterns/shapes is possible.

In summary, high GATB dimension correlations for actual test score and regression-estimated data suggest these scores reflect little more than *g*. As such, they essentially only allow for occupational differentiation and grouping according to differences in *g* levels. In contrast, although nontrivial, GATB dimension correlations for

79

analyst data are much lower, suggesting the scores reflect more than *g*. These data may then allow for differentiation and grouping according to differences in both ability level and shape. These characteristics of the data not only qualify the results of this study, but also suggest that analyst data may be more appropriate than test score data when the goal is to differentiate among or group occupations according to ability patterns/shapes (e.g., for matching individuals to occupations according to strengths and weaknesses). The following sections describe the focal findings of this study: the clustering results and their implications.

Clustering Results

Two general conclusions seem reasonable based on the clustering results. First, there does not appear to be a clear number of clusters underlying any of the datasets used. That is, no clear solutions in terms of number of clusters were revealed for actual test score, analyst, or regression-estimated data at the DOT and SOC level. The reason for this is unclear. It may be that no clear cluster structure underlies the data. For example, aptitude requirements for the occupations included in this study may be distributed relatively evenly or continuously, rather than in a disjointed or grouped manner. Although clusters could be created in this situation, the number of clusters would obviously be hard to identify and the groupings would be fairly artificial (rather than being reflective of the true underlying structure of the data). Without knowing the underlying structure of the data, it is difficult to determine the likelihood of this possibility. However, it should be noted that this type of situation may not be all that uncommon in practice, where clusters are formed whether a clear structure appears to exist or not. Thus, the analyses and results of this study are relevant despite this potential

limitation; they represent an attempt to do the best one can empirically, which may be fairly common in practice.

A related possibility is that a 'true' cluster structure does exist but the several potential sources of error in estimating occupational ability requirements discussed previously (e.g., cognitive limitations and biases, sampling error) acted to obscure this underlying structure or pattern in data used in this study. As one example, halo error in analyst ratings may have blurred any clear differences among groups. Similar problems may have existed in the other data types as well. This obscuring of the 'true' cluster structure would then have made it difficult to identify an appropriate number of clusters.

Alternatively, it may be that a reasonably clear cluster structure underlies the data, but the methods used in this study did not detect the number of clusters. For example, although a relatively wide range of cluster solutions were examined (solutions ranging from 2 to 50 clusters for each type of data at both the DOT and SOC level) the true number of clusters might exist beyond this range. Although perhaps less practically useful, the true number of clusters underlying each dataset may be greater than 50. In addition, the indices used to detect the number of clusters could have been ineffective. Although previous research (Milligan & Cooper, 1985) has indicated the three indices used in this study are among the best available, it is difficult to determine how effective an index will be in a given situation.

In sum, it is not apparent whether the absence of a clear number of clusters in the data is due to the underlying structure of the data or the methods used to detect the number of clusters. However, given the relatively wide range of cluster solutions examined and research indicating the quality of indices used, it may be more likely that

these results are a function of the underlying data structure. In any case, multiple cluster solutions were used for subsequent analyses to allow for comparisons across cluster solutions containing different numbers of clusters and avoid basing conclusions on one questionable cluster solution for each type of data.

The second general conclusion that can be drawn from the clustering results is that the three types of ability data produce substantially different cluster solutions. Across data types, cluster ranges, and at both the DOT and SOC level there appeared to be only minimal cluster solution agreement above chance levels. It appears that the differences in the source of information, process of development, and potential sources of bias/inaccuracy across the three types of data resulted in differing estimates of occupational aptitude requirements to some extent, which in turn resulted in differing occupational clusters. Although analyst and regression-estimated data tended to produce relatively more similar solutions than the other two data type combinations (i.e., actual test score and analyst data and actual test score and regression-estimated data), this combination's level of agreement was still quite low. Thus, it appears that these three data types produce dissimilar occupational groupings, with actual test score data producing the most dissimilar solutions.

This finding extends previous research indicating that different types of occupational data often result in substantially different job clusters. Specifically, prior research suggested that job characteristic data from *different* psychological domains can result in different job clusters (Pearlman, 1980; Cornelius, Carron, & Collins, 1979). The present study indicates that different types of data from within the *same* psychological domain (abilities) can produce considerably different clusters as well. This may have

implications for the development of job characteristic data to be used for job clustering. In particular, previous findings suggested that choosing the psychological domain according to the purpose of clustering is essential when developing data for clustering. The current findings extend this, suggesting that choosing the type of data (or how the data is developed) within a given psychological domain may also be important. That is, because both the choice of psychological domain and type of data within a given domain can substantially influence job clustering results, it may be important to consider both when developing data to be used in job clustering. The following sections discuss this further.

Implications for Career Exploration/O*NET's Ability Profiler

As noted previously, O*NET (the Department of Labor's computerized occupational information tool developed to replace the *Dictionary of Occupational Titles*) includes a career exploration tool called the "Ability Profiler." During the development of this tool, developers generated regression-estimated ability scores for each DOT occupation. These estimated scores were developed using regression weights obtained from the prediction of mean incumbent GATB test scores from data available from the DOT. Therefore, the regression-estimated scores represent predicted mean incumbent GATB test scores.

However, clustering results from this study suggest that mean incumbent GATB test score profiles (i.e., actual test score data) may be substantially different from regression-estimated ability profiles. As discussed above, cluster solutions resulting from actual test score data demonstrated little similarity to cluster solutions resulting from regression-estimated data. This indicates that many of the occupations considered similar

(i.e., belonging to the same cluster) when described by actual test score data are considered dissimilar (i.e., belong to different clusters) when described by regression-estimated data. This suggests that actual test score profiles and regression-estimated profiles may be considerably dissimilar, and therefore the "Ability Profiler" might function differently if it included actual test score data (those data the regression-estimated scores are supposed to predict) rather than regression-estimated data. Specifically, the Profiler might produce differing occupational suggestions for job seekers depending upon what type of occupational data is being used. Although this does not necessarily mean that the Profiler is producing inappropriate occupational suggestions, it does indicate that the recommendations currently produced by this tool may be data dependent to some extent. In other words, the same occupations might not be suggested to a given individual if another type of occupational ability data were used.

Implications for Test Validation

Several general patterns emerged from the criterion-related validity results in terms of the level, flatness, and shape of occupational validity profiles. First, whereas actual test score data tended to produce clusters that differed significantly in terms of overall profile level, analyst and regression-estimated data did not. At both the DOT and SOC level and for all three cluster ranges, actual test score data produced significantly different validity profiles in terms of overall level. However, analyst data failed to produce significantly different validity profiles in any of the six analyses. Furthermore, regression-estimated data resulted in significantly different validity profiles in only one out of the six analyses. Partial eta squared values demonstrated a similar pattern,

indicating that cluster membership accounted for more variance in mean validity for actual test score data than for analyst and regression-estimated data.

Thus, this seems to indicate that actual test score data tend to perform better than analyst or regression-estimated in terms of producing clusters that differ in overall validity profile level. However, although this pattern was consistent across all analyses, the differences were relatively small and should not be over interpreted. In addition, the effect sizes obtained for this effect (ranging from approximately 0.02 to 0.20) were relatively small, indicating the presence of nontrivial amounts of within cluster variance. This suggests that cluster solutions resulting from all three types of data may be less that ideal for use in test validation purposes. Furthermore, significant shape differences (discussed below) qualify the level results to some extent.

Second, the flatness analyses clearly demonstrated significant differences in mean criterion-related validities across GATB scales. Specifically, results indicated that for the occupations included in this study, G and N tended to be the best predictors of performance, P and Q were moderate predictors, and V, S, K, F, and M were least predictive. For the most part, these findings are not surprising. For example, measures of general cognitive ability (G) have been shown to be predictive of performance in a wide variety of jobs (e.g., Hunter, 1983a; 1983b; Hunter & Hunter, 1984; Schmidt, Hunter, & Pearlman, 1980). Thus, this variable should be a relatively good predictor of performance in most of the occupations included in this study, resulting in a high overall mean validity coefficient. On the other hand, abilities such as motor coordination (K) and finger dexterity (F) are likely to be necessary for only a relatively smaller subset of jobs (e.g., those requiring a considerable amount of physical activity). Therefore, validity

coefficients for these abilities should be more variable across jobs, resulting in relatively lower overall mean coefficients.

The one somewhat unusual result obtained from these analyses involves verbal ability (V). Although G and N tended to be the best predictors of performance, V appeared to be among the worst. This is somewhat surprising in that G, N, and V all measure aspects of cognitive ability, and although they obviously measure different constructs, all three should quite g-loaded. However, only G and N were good predictors. The cause of this result is unclear but it may be at least partly a function of the sample of occupations in this study. For example, this sample may contain a disproportionate amount of occupations that do in fact require cognitive ability (G), but are more oriented towards numerical/mathematical work (N) rather than verbal tasks (V). Therefore, on average G and N are better predictors of performance than is V.

Third, parallelism tests indicated that all three types of data produced clusters that differ significantly in terms of the shape of their criterion-related validity profiles. At both the DOT and SOC level and for all three cluster ranges, each type of ability data produced validity profiles of differing shape. In addition, partial eta squared values indicated that at the DOT level, shape differences in regression-estimated clusters accounted for more variance than did shape differences in analyst clusters; shape differences in analyst clusters, in turn, accounted for more variance than did shape differences in actual test score clusters. However, this pattern did not hold at the SOC level. In fact, no clear pattern emerged from SOC-level parallelism analyses. Therefore, it is difficult to draw conclusions regarding the relative merits of each type of data in terms of between-to-within cluster variability in validity profiles.

This lack of a clear conclusion may reflect the fact that, generally speaking, each type of data has its own strengths and weaknesses, but overall there is no strong reason to believe that one data type has superior qualities, particularly with respect to test validation purposes. As noted in the discussion of the source and process involved in the development of the three data types, although each may have unique strengths, they also are likely to have unique weaknesses. For example, although analysts are able to be more flexible than the regression equation in rating jobs, with this flexibility comes less reliability.

More importantly, none of these unique strengths and weaknesses appear to be directly relevant to test validation situations. That is, there does not seem to be any convincing reason, based on conceptual analysis of the three data types, to assume that one type will necessarily be more useful for test validation. Thus, a priori, there appeared to be reason to believe that clusters resulting from these three data types would be fairly dissimilar. However, there were not compelling reasons to believe one data type would be the most useful in test validation situations. Results seem to parallel this a priori reasoning: the three data types produced dissimilar cluster solutions, but none appeared to be superior for test validation.

In addition, although significant results were obtained for all parallelism analyses, shape differences accounted for only small to moderate amounts of variance. For example, partial eta squared values varied from 0.05 to 0.26 for these effects. Although in some contexts these values may be acceptable, they could be viewed as relatively small for the present purposes. Specifically, these values indicate the presence of substantial amounts of within cluster variability in validity profile shapes. If these job

clusters were used for test validation purposes, these considerable differences in predictor-criterion relationships would be masked. This type of situation might then lead to incorrect conclusions in the validation process, such as concluding that a predictor is not valid for all the jobs in a cluster, when in fact it is valid for some jobs in that cluster, or that a predictor is valid for all the jobs in a cluster, when in fact it is not valid for some jobs in that cluster. Therefore, not only is it difficult to draw firm conclusions regarding the relative merits of the three types of data for test validation purposes, it appears that using clusters resulting from any of the data types could lead to inappropriate conclusions in test validation situations.

This outcome may at least partly reflect the difficulty encountered in finding clear cluster solutions. As discussed in the results, no clear cluster solutions emerged for any of the data sets, even when examining multiple cluster ranges. This may be an indication that a clear cluster structure did not underlie any of the datasets, and therefore, any cluster solutions obtained from these data sets may have been relatively "artificial" (i.e., the solutions do not reflect the underlying data structure). The forced nature of these cluster solutions may then make them ineffective for most purposes, as numerous occupations are likely to have been grouped and separated inappropriately. Again, however, situations where a cluster structure must be 'forced' may not be that uncommon in practice.

Finally, it should be noted that these criterion-related validity results may be complicated to some extent by range restriction effects. Some of the variability in criterion-related validity coefficients might have been due to differential range restriction across jobs, rather than true differences in predictor-criterion relationships. For example,

the range of cognitive ability may be relatively more restricted in samples of incumbents from higher level jobs than in those from entry level jobs. This differential restriction in range might then result in different estimates of criterion-related validity for cognitive aptitudes (e.g., G, V, N) for high versus low level jobs, even if the true predictor-criterion relationship is similar. The possibility of the presence of these effects must be considered when interpreting these results.

Summary and Implications

In sum, results indicate that actual test score data tend to produce clusters that differ significantly in terms of overall validity profile levels, whereas analyst and regression-estimated data do not. In addition, the mean validity profile across clusters was not flat; G and N tended to be better predictors than the other GATB scales. However, these findings are qualified to some extent by results indicating that each data type produced clusters differing significantly in validity profile shape. However, effect sizes obtained from both the level and shape analyses were quite small, indicating the presence of large amounts of within cluster variability in criterion-related validity.

Overall, these results do not indicate the superiority of one type of data for use in job clustering for test evaluation purposes. Although the level analyses seemed to indicate that actual test score data consistently performed better, this did not hold for the shape analyses. More importantly, the magnitude of effect sizes for level and shape effects seemed to indicate that the job clusters produced in this study might not be effective for test validation purposes, regardless of which data type was used. These findings highlight the importance of carefully examining job clusters before putting them to use. Cluster analyses performed for this study failed to reveal a clear cluster solution

for any of the data types. This may have been an early sign that any cluster solutions obtained from these datasets would be rather artificial and therefore less than ideal for most purposes.

Implications for Job Evaluation

Overall, pay rate results failed to demonstrate any meaningful differences across data types in terms of their usefulness in job evaluation. Confidence intervals for the three data types overlapped substantially in all three cluster ranges. These finding suggest the three data types are equally effective for use in job evaluation situations. Not unexpectedly, cluster solutions consisting of a larger number of clusters (e.g., those in the 35-54 cluster range) tended to produce higher ICCs than solutions consisting of fewer clusters (e.g., those in the 2-14 cluster range). Finally, generally speaking, cluster membership accounted for moderate amounts of pay rate variability. ICCs ranged from 0.286 to 0.657, indicating the presence of nontrivial amounts of within cluster variance. This suggests caution should be used if ability based job clusters are to be used during the job evaluation process.

Failure to find meaningful differences across data types in terms of performance for job evaluation purposes further emphasizes the notion that none of the three data types necessarily has superior qualities or is more useful overall. Each data type has strengths and weaknesses that affect its performance in job clustering situations. More importantly, these strengths and weaknesses do not appear to be directly relevant to job evaluation, as all three data types performed similarly with respect to pay rates.

Overall, cluster membership accounted for a reasonable amount of variance in pay rates for all three data types. However, a substantial amount of within cluster variability

remained in all cluster solutions. This suggests ability-based job clusters may be useful in job evaluation situations, but clearly these groupings should only be used as one part of the larger evaluation process. For example, broad ability clusters could be used to categorize and evaluate jobs initially. Then, other data (e.g., tasks performed) and considerations could be used to further categorize and evaluate jobs to establish appropriate pay rates. In any case, it appears that ability-based clusters could be a useful part of the job evaluation process.

Summary and Implications

In sum, pay rate analyses failed to reveal significant differences in performance across data types, although differences were found across cluster solution ranges. Overall, cluster membership accounted for a moderate amount of variance in pay rates. Thus, ability-based job clustering could be a useful part of the overall job evaluation process. Furthermore, this study suggests that type of ability data chosen for this purpose may not have meaningful effects on the outcome of the evaluation process.

Conclusions

The purpose of this study was to examine similarities and differences in job clusters produced by three types of ability data: actual test score, analyst, and regression estimated. Results indicated that these three data types produced substantially different job clusters. However, these differences did not appear to have clear implications for test validation or job evaluation. It appears that the manner in which job characteristics data are developed can have an important influence on job clustering. However, this influence may not always have a clear effect on personnel-related functions using these clusters. Overall, these findings complement previous research indicating that job characteristic

data from different psychological domains can produce substantially different job clusters by demonstrating a similar effect for different types of job characteristic data from within the same domain. Combined, these findings suggest that both the methods used to develop job characteristics data and the psychological domain to which they belong can have an important influence on job clustering.

REFERENCES

Arvey, R. D., & Mossholder, K. M. (1977). A proposed methodology for determining similarities and differences among jobs. *Personnel Psychology, 30*, 363-374.

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43*, 695-716.

Bemis, S. E. (1968). Occupational validity of the General Aptitude Test Battery. *Journal of Applied Psychology, 52*, 240-244.

Bureau of Labor Statistics. (2000). http://stats.bls.gov/blshome/htm.

Bureau of Labor Statistics. (2001). http://www.bls.gov/oes/2000/oestec2000.htm.

Cain, P. S., & Green, B. F. (1983). Reliabilities of selected ratings available from the Dictionary of Occupational Titles. *Journal of Applied Psychology, 68*, 155-165.

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*, 1-27.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge, UK: Cambridge University Press.

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationship between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82*, 300-310.

Colihan, J., & Burger, G. K. (1995). Constructing job families: An analysis of quantitative techniques used for grouping jobs. *Personnel Psychology, 48*, 563-586.

Cooper, W. H. (1981). Ubiquitous Halo. *Psychological Bulletin, 90*, 218-244.

Cornelius, E. T., Carron, T. J., & Collins, M. N. (1979). Job analysis models and job classification. *Personnel Psychology, 32*, 693-708.

Cornelius, E. T., Hakel, M. D., & Sackett, P. R. (1979). A methodological approach to job classification for performance appraisal purposes. *Personnel Psychology, 32*, 283-297.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668-1674.

Desmarais, L. B., & Sackett, P. R. (1993). Investigating a cognitive complexity hierarchy of jobs. *Journal of Vocational Behavior, 43*, 279-297.

Desmond, R. E., & Weiss, D. J. (1973). Supervisor estimation of abilities required in jobs. *Journal of Vocational Behavior, 3*, 181-194.

Desmond, R. E., & Weiss, D. J. (1975). Worker estimation of ability requirements of their jobs. *Journal of Vocational Behavior, 7*, 13-27.

Donner A., & Wells G. (1986). A comparison of confidence interval methods for the intraclass correlation coefficient. *Biometrics, 42*, 401-412.

Droege, R. C. (1968). GATB longitudinal validation study. *Journal of Counseling Psychology, 15*, 41-47.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: Wiley.

Dye, D., & Silver, M. (1999). The origins of O*NET. In N. G. Peterson & M. D. Mumford (Eds.), *An occupational information system for the 21st century: The development of O*NET* (pp. 9-19). Washington, DC: American Psychological Association.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66*, 127-148.

Fischer, D. G., & Sobkow, J. (1979). Workers' estimation of ability requirements of their jobs. *Perceptual and Motor Skills, 48*, 519-531.

Geyer, P. D., Hice, J., Hawk, J., Boese, R., & Brannon, Y. (1989). Reliabilities of ratings available from the Dictionary of Occupational Titles. *Personnel Psychology, 42*, 547-560.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests*. New York: Wiley.

Gottfredson, L. S. (1986). Occupational aptitude patterns map: Development and implications for a theory of job aptitude requirements. *Journal of Vocational Behavior, 29*, 254-291.

Hakistian, A. R., & Bennet, R. W. (1978). Validity studies using the Comprehensive Ability Battery (CAB): II. Relationships with the DAT and GATB. *Educational and Psychological Measurement, 38*, 1003-1015.

Harris, R. J. (1975). *A primer of multivariate statistics*. New York: Academic Press.

Hartman, E. A., Mumford, M. D., & Mueller, S. (1992). Validity of job classifications: An examination of alternative indictors. *Human Performance, 5*, 191-211.

Harvey, R. J. (1986). Quantitative approaches to job classification: A review and critique. *Personnel Psychology, 39*, 267-289.

Holt, R. R. (1970). Yet another look at clinical and statistical prediction: Or, is clinical psychology worthwhile? *American Psychologist, 25*, 337-349.

Huber, V. L. (1991). Comparison of supervisor-incumbent and female-male multidimensional job evaluation ratings. *Journal of Applied Psychology, 76*, 115-121.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193-218.

Hunter, J. E. (1983a). *The dimensionality of the General Aptitude Test Battery (GATB) and the dominance of general factors over specific factors in the prediction of job performance for U.S. Employment Service* (Test Research Rep. No. 44). Washington, DC: U.S. Department of Labor.

Hunter, J. E. (1983b). *Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery (GATB)* (Test Research Rep. No. 45). Washington, DC: U.S. Department of Labor.

Hunter, J. E. & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Knapp, R. P., Knapp, L., & Michael, W. B. (1977). Stability and concurrent validity of the Career Ability Placement Survey (CAPS) against the DAT and the GATB. *Educational and Psychological Measurement, 37*, 1081-1085.

Krzystofiak, F., Cardy, R., & Newman, J. (1988). Implicit personality and performance appraisal: The influence of trait inferences on evaluations of behavior. *Journal of Applied Psychology, 73*, 515-521.

Marchese, M. C. (1992). Clinical versus actuarial prediction: A review of the literature. *Perceptual and Motor Skills, 75*, 583-594.

McCloy, R. A., Campbell, J. P., & Oswald, F. L. (1999). *Generation and use of occupation ability profiles for exploring O*NET occupations*. Alexandria, VA: Human Resources Research Organization.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Milkovich, G. T., & Newman, J. M. (1990). *Compensation*. Homewood, IL: BPI Irwin.

Miller, A. R., Treiman, D. J., Cain, P. S., Roos, P. A. (Eds.). (1980). *Work, jobs, and occupations: A critical review of the Dictionary of Occupational Titles*. Washington, DC: National Academy Press.

Milligan, G. W. (1981). A review of Monte Carlo tests of cluster analysis. *Multivariate Behavioral Research, 16*, 379-407.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*, 159-179.

Milligan, G. W., & Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research, 21*, 441-458.

Milligan, G. W., & Cooper, M. C. (1987). Methodology review: Clustering methods. *Applied Psychological Measurement, 11*, 329-354.

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology, 82*, 627-655.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage Publications.

Murphy, K. R., & Davidshofer, C. O. (2001). *Psychological testing: Principles and applications*. Upper Saddle River, NJ: Prentice Hall.

Murphy, K. R., & Jako, R. (1989). Under what conditions are observed intercorrelations greater or smaller than true intercorrelations? *Journal of Applied Psychology, 74*, 827-830.

Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology, 78*, 218-225.

Murphy, K. R., & Reynolds, D. H. (1988). Does true halo affect observed halo? *Journal of Applied Psychology, 73*, 235-238.

National Crosswalk Service Center. (2001). http://www.state.ia.us/ncdc/index.html.

O'Reilly, A. P. (1973). Skill requirements: Supervisor-subordinate conflict. *Personnel Psychology, 26*, 75-80.

Passini, F. T., & Norman, W. T. (1966). A universal conception of personality structure? *Journal of Personality and Social Psychology, 4,* 44-49.

Pearlman, K. (1980). Job families: A review and discussion of their implications for personnel selection. *Psychological Bulletin, 87,* 1-28.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R. & Fleishman, E. (Eds.). (1999). *An occupational information system for the 21st century: The development of O*NET.* Washington, DC: American Psychological Association.

Pulakos, E. D., Schmitt, N., & Ostroff, C. (1986). A warning about the use of a standard deviation across dimensions within ratees to measure halo. *Journal of Applied Psychology, 71,* 29-32.

Rand, W. M. (1971). Objective criteria for he evaluation of clustering methods. *Journal of the American Statistical Association, 66,* 846-850.

Sarle, W. S. (1983). *Cubic clustering criterion* (SAS Tech. Rep. No. A-108). Cary, NC: SAS Institute Inc.

SAS Institute (1999). *SAS/STAT user's guide, version 8.* Cary, NC: Author.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66,* 178-200.

Schmidt, F. L., Hunter, J. E. & Pearlman, K. (1980). Task difference and validity of aptitude tests in selection: A red herring. *Journal of Applied Psychology, 66,* 166-185.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Smith, J. E., & Hakel, M. D. (1979). Convergence among data sources, response bias, and reliability and validity of a structured job analysis questionnaire. *Personnel Psychology, 32,* 677-692.

Spaeth, J. L. (1979). Vertical differentiation among occupations. *American Sociological Review, 44,* 746-762.

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

Steele, C. M. (1997). A threat in the air. *American Psychologist, 52,* 613-629.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797-811.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston: Allyn and Bacon.

U. S. Department of Labor. (1970). *Manual for the USTES General Aptitude Test Battery.* Washington, DC: U.S. Government Printing Office.

U.S. Department of Labor. (1979). *Manual for the USES General Aptitude Test Battery.* Washington, DC: U.S. Government Printing Office.

U.S. Department of Labor. (1991). *The revised handbook for analyzing jobs.* Washington, DC: U.S. Government Printing Office.

U.S. Department of Labor. (1998). *O*NET 98 viewer: User's guide for version 1.0.* Washington, DC: U.S. Government Printing Office.

Waldman, D. A., Yammarino, F. J., & Avolio, B. J. (1990). A multiple level investigation of personnel ratings. *Personnel Psychology, 43,* 811-835.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58,* 236-244.

Wilk, S. L., Desmarais, L. B., & Sackett, P. R. (1995). Gravitation to jobs commensurate with ability: Longitudinal and cross-sectional tests. *Journal of Applied Psychology, 80,* 79-85.

Wilk, S. L., & Sackett, P. R. (1996). Longitudinal analysis of ability-job complexity fit and job change. *Personnel Psychology, 49,* 937-967.

[1] Two of the standard nine GATB aptitudes, General Intelligence and Numerical Ability, are not included because General Intelligence is simply a composite of Vocabulary and Arithmetic Reasoning and thus is excluded from Ability Profiler analyses; Numerical Ability has been split into its two component tests (Arithmetic Reasoning and Computation; McCloy, Campbell, & Oswald, 1999).

Table 1

48 DOT Variables Used to Predict GATB Scores

Data, People, Things

Reasoning, Math, Language

Specific Vocational Preparation

Physical Demands: Strength, Climbing, Balance, Stooping, Kneeling, Crouching,

Crawling, Reaching, Handling, Fingering, Feeling, Talking, Hearing,

Tasting/Smelling, Near Acuity, Far Acuity, Depth Perception, Accommodation, Color

Vision, Field of Vision

Temperaments: Directing, Repetitive, Influencing, Variety, Expressing Stress,

Tolerances, Under, People, Judgments

GATB Aptitude Ratings: G – General Intelligence, V – Verbal Ability, N – Numerical

Ability, S – Spatial Ability, P – Form Perception, Q – Clerical Ability, K – Motor

Coordination, F – Finger Dexterity, M – Manual Dexterity, E – Eye-Hand-Foot

Coordination, C – Color Discrimination

Table 2

Sources of Potential Inaccuracy in Job Analysis as Described by Morgeson and Campion

(1997)

Social Sources	Cognitive Sources
(1) Social Influence Processes	(1) Limitations in Information Processing Systems
• Conformity Pressures	• Information Overload
• Extremity Shifts	• Heuristics
• Motivation Loss	• Categorization
(2) Self-Presentation Processes	(2) Bias in Information Processing Systems
• Impression Management	• Carelessness
• Social Desirability	• Extraneous Information
• Demand Effects	• Inadequate Information
	• Order and Contrast Effects
	• Halo
	• Leniency and Severity
	• Method Effects

Table 3

Criteria for Evaluating Job Clusters

Objectives of Job Clustering	Ability Data Used for This Objective?	Criteria Available for This Objective for the Present Sample?
Test validation (personnel selection) (Arvey & Mossholder, 1977)	Yes	Yes
Job evaluation (for setting pay structures, wage and salary administration) (Pearlman, 1980)	Yes	Yes
Vocational and educational guidance (Pearlman, 1980)	Yes	No
Job placement (Pearlman, 1980)	Yes	No
Personnel classification (Pearlman, 1980)	Yes	No
Establishing career promotion ladders (career-path planning) and lines of transfer (Pearlman, 1980)	Yes	No
Internal job classification (Pearlman, 1980)	Yes	No

Table 3 (cont'd).

	Yes	No
Exploratory research, theory development, and methodological research objectives (Pearlman, 1980)	Yes	No
Performance appraisal (e.g., Cornelius, Hakel, & Sackett, 1979)	No	No
Establishment of vocational training curricula (Pearlman, 1980)	No	No
Development of training programs (Pearlman, 1980)	No	No
Population-level occupational data collection and analysis for economic and social purposes (Pearlman, 1980)	No	No

Table 4

Number of SOCs in Dataset per SOC Major Group

SOC Major Group	Description	Number of SOCs in Dataset
11-0000	Management Occupations	6
13-0000	Business and Financial Operations Occupations	8
15-0000	Computer and Mathematical Occupations	5
17-0000	Architecture and Engineering Occupations	12
19-0000	Life, Physical, and Social Science Occupations	8
21-0000	Community and Social Services Occupations	5
23-0000	Legal Occupations	3
25-0000	Education, Training, and Library Occupations	4
27-0000	Arts, Design, Entertainment, Sports, and Media Occupations	4
29-0000	Healthcare Practitioners and Technical Occupations	18
31-0000	Healthcare Support Occupations	5
33-0000	Protective Service Occupations	8

Table 4 (cont'd).

35-0000	Food Preparation and Serving Related Occupations	8
37-0000	Building and Grounds Cleaning and Maintenance Occupations	2
39-0000	Personal Care and Service Occupations	7
41-0000	Sales and Related Occupations	9
43-0000	Office and Administrative Support Occupations	27
45-0000	Farming, Fishing, and Forestry Occupations	2
47-0000	Construction and Extraction Occupations	20
49-0000	Installation, Maintenance, and Repair Occupations	22
51-0000	Production Occupations	68
53-0000	Transportation and Material Moving Occupations	13
55-0000	Military Specific Occupations	0

Table 5

Means, Standard Deviations, and Intercorrelations: Actual Test Score Data (DOT Level)

GATB Aptitude	M	SD	1	2	3	4	5	6	7	8
1. Verbal Ability	98.85	10.72	--							
2. Numerical Ability	97.97	12.19	0.90	--						
3. Spatial Ability	102.42	10.93	0.76	0.84	--					
4. Form Perception	103.89	10.86	0.77	0.78	0.78	--				
5. Clerical Ability	105.93	11.45	0.81	0.72	0.56	0.85	--			
6. Motor Coordination	102.39	9.16	0.72	0.64	0.46	0.79	0.87	--		
7. Finger Dexterity	97.94	8.05	0.43	0.46	0.46	0.61	0.42	0.58	--	
8. Manual Dexterity	105.13	8.17	0.32	0.39	0.45	0.57	0.43	0.53	0.60	--

Note. N = 518.

All correlations are significant at *p* < .01.

Table 6

Means, Standard Deviations, and Intercorrelations: Analyst Data (DOT Level)

GATB Aptitude	M	SD	1	2	3	4	5	6	7	8
1. Verbal Ability	3.21	.80	--							
2. Numerical Ability	3.38	.81	0.71**	--						
3. Spatial Ability	3.30	.82	0.35**	0.48**	--					
4. Form Perception	3.23	.74	0.26**	0.40**	0.70**	--				
5. Clerical Ability	3.61	.83	0.63**	0.59**	0.13**	0.16**	--			
6. Motor Coordination	3.28	.60	-0.08	0.00	0.35**	0.40**	0.01	--		
7. Finger Dexterity	3.30	.65	0.00	0.10*	0.37**	0.46**	0.05	0.62**	--	
8. Manual Dexterity	3.14	.58	-0.21**	-0.10*	0.38**	0.34**	-0.24**	0.56***	0.50**	--

Note. N = 518.

*$p < .05$. **$p < .01$.

Table 7

Means, Standard Deviations, and Intercorrelations: Regression Estimated Data (DOT Level)

GATB Aptitude	M	SD	1	2	3	4	5	6	7	8
1. Verbal Ability	98.95	8.68	--							
2. Numerical Ability	105.99	18.90	0.97	--						
3. Spatial Ability	102.65	8.43	0.84	0.94	--					
4. Form Perception	104.08	7.10	0.97	0.98	0.90	--				
5. Clerical Ability	106.06	8.02	0.97	0.92	0.74	0.95	--			
6. Motor Coordination	102.60	5.88	0.94	0.85	0.64	0.90	0.97	--		
7. Finger Dexterity	98.15	3.10	0.79	0.77	0.74	0.86	0.77	0.81	--	
8. Manual Dexterity	105.21	2.47	0.75	0.81	0.88	0.86	0.70	0.67	0.91	--

Note. $N = 518$.

All correlations are significant at $p < .01$.

Table 8

Means, Standard Deviations, and Intercorrelations: Actual Test Score Data (SOC Level)

GATB Aptitude	M	SD	1	2	3	4	5	6	7	8
1. Verbal Ability	102.25	10.82	--							
2. Numerical Ability	101.38	11.29	0.89	--						
3. Spatial Ability	105.59	10.05	0.70	0.81	--					
4. Form Perception	106.51	10.34	0.78	0.80	0.76	--				
5. Clerical Ability	108.87	11.21	0.84	0.74	0.53	0.87	--			
6. Motor Coordination	104.23	8.53	0.77	0.66	0.45	0.79	0.88	--		
7. Finger Dexterity	98.20	7.51	0.51	0.55	0.57	0.69	0.51	0.62	--	
8. Manual Dexterity	105.40	7.44	0.34	0.44	0.52	0.57	0.43	0.50	0.55	--

Note. N = 264.

All correlations are significant at *p* < .01.

Table 9

Means, Standard Deviations, and Intercorrelations: Analyst Data (SOC Level)

GATB Aptitude	M	SD	1	2	3	4	5	6	7	8
1. Verbal Ability	2.94	.81	--							
2. Numerical Ability	3.16	.76	0.65**	--						
3. Spatial Ability	3.15	.84	0.19**	0.39**	--					
4. Form Perception	3.14	.76	0.15*	0.37**	0.72**	--				
5. Clerical Ability	3.40	.76	0.56**	0.47**	-0.10	-0.03	--			
6. Motor Coordination	3.28	.60	-0.21**	-0.08	0.40**	0.43**	-0.10	--		
7. Finger Dexterity	3.31	.66	-0.04	0.08	0.41**	0.51**	0.00	0.68**	--	
8. Manual Dexterity	3.17	.63	-0.30**	-0.19**	0.42**	0.42**	-0.35**	0.67**	0.61**	--

Note. N = 264.

*$p < .05$. **$p < .01$

110

Table 10

Means, Standard Deviations, and Intercorrelations: Regression Estimated Data (SOC Level)

GATB Aptitude	M	SD	1	2	3	4	5	6	7	8
1. Verbal Ability	102.16	8.26	--							
2. Numerical Ability	112.67	17.15	0.96	--						
3. Spatial Ability	105.35	7.36	0.77	0.89	--					
4. Form Perception	106.47	6.67	0.97	0.97	0.85	--				
5. Clerical Ability	108.80	8.00	0.97	0.91	0.64	0.94	--			
6. Motor Coordination	104.50	6.16	0.94	0.83	0.55	0.90	0.97	--		
7. Finger Dexterity	99.04	3.18	0.73	0.70	0.65	0.82	0.71	0.78	--	
8. Manual Dexterity	105.88	2.33	0.65	0.72	0.83	0.80	0.59	0.59	0.89	--

Note. N = 264.

All correlations are significant at $p < .01$.

Table 11

Reliability Estimates from Geyer, et al. (1989)

GATB Aptitude	Alpha: One Rater	Alpha: Four Raters
G	0.88	0.97
V	0.93	0.98
N	0.75	0.92
S	0.78	0.93
P	0.72	0.91
Q	0.70	0.90
K	0.65	0.88
F	0.68	0.89
M	0.51	0.81

Table 12

Number of Clusters Indicated by the CCC, Pseudo F, and Pseudo r^2

Data Type	Small (2-14 Clusters)	Medium (15-34 Clusters)	Large (35-54 Clusters)
Actual Test Score (DOT)	3	18	42
Analyst (DOT)	3	23	48
Regression Estimated (DOT)	3	23	50
Actual Test Score (SOC)	3	26	39
Analyst (SOC)	3	21	40
Regression Estimated (SOC)	4	22	42

Table 13

Adjusted Rand Statistic: DOT Level

Comparison	2-14 Cluster Range	15-34 Cluster Range	35-54 Cluster Range
Actual Test Score and Analyst	0.1867	0.0683	0.0346
Analyst and Regression-Estimated	0.4166	0.1546	0.1175
Actual Test Score and Regression-Estimated	0.1945	0.0607	0.0328

Note. Rand adjustment from Hubert and Arabie (1985).

Table 14

Adjusted Rand Statistic: SOC Level

Comparison	2-14 Cluster Range	15-34 Cluster Range	35-54 Cluster Range
Actual Test Score and Analyst	0.1121	0.0773	0.0662
Analyst and Regression-Estimated	0.1539	0.1779	0.1573
Actual Test Score and Regression-Estimated	0.2376	0.0700	0.0572

Note. Rand adjustment from Hubert and Arabie (1985).

Table 15

Criterion-Related Validity Study Sample Size Means and Standard Deviations

GATB Aptitude	Mean	SD
G	91.37	98.470
V	92.17	87.576
N	92.17	87.576
S	92.17	87.576
P	18.55	16.942
Q	18.37	17.265
K	16.15	16.233
F	16.01	17.152
M	17.03	18.415

Table 16

Criterion Related Validity Coefficient Descriptive Statistics: DOT Level

GATB Aptitude	Mean	Std. Error of Mean	Median	Std. Deviation	Minimum	Maximum
G	.233	.0072	.240	.1645	-.81	.78
V	.166	.0072	.166	.1646	-.83	.69
N	.221	.0070	.220	.1599	-.81	.88
S	.158	.0071	.165	.1637	-.61	.65
P	.184	.0071	.180	.1637	-.55	.71
Q	.186	.0071	.191	.1629	-.48	.80
K	.154	.0068	.152	.1553	-.62	.58
F	.153	.0075	.155	.1728	-.52	.69
M	.160	.0081	.155	.1846	-.87	.66

Note. $N = 518$ for each GATB Aptitude.

Table 17

Profile Analysis "Levels" Test: 2-14 Cluster Range, DOT Level

	Source	df	F	Sig.	Partial Eta Squared
Actual Test Score	Intercept	1	1097.324	.000	.681
	3 Clusters	2	4.484	.012	.017
	Error	515			
Analyst	Intercept	1	1130.666	.000	.687
	3 Clusters	2	.674	.510	.003
	Error	515			
Regression-Estimated	Intercept	1	1122.975	.000	.686
	3 Clusters	2	1.183	.307	.005
	Error	515			

Table 18

Profile Analysis "Levels" Test: 15-34 Cluster Range, DOT Level

	Source	df	F	Sig.	Partial Eta Squared
Actual Test Score	Intercept	1	848.718	.000	.629
	18 Clusters	17	3.324	.000	.102
	Error	500			
Analyst	Intercept	1	815.802	.000	.622
	23 Clusters	22	.974	.496	.041
	Error	495			
Regression-Estimated	Intercept	1	978.309	.000	.664
	23 Clusters	22	1.787	.016	.074
	Error	495			

Table 19

Profile Analysis "Levels" Test: 35-54 Cluster Range, DOT Level

	Source	df	F	Sig.	Partial Eta Squared
Actual Test Score	Intercept	1	787.327	.000	.623
	42 Clusters	41	2.832	.000	.196
	Error	476			
Analyst	Intercept	1	771.542	.000	.621
	48 Clusters	47	.833	.777	.077
	Error	470			
Regression-Estimated	Intercept	1	843.382	.000	.643
	50 Clusters	49	1.135	.254	.106
	Error	468			

Table 20

Profile Analysis "Flatness" Test: 2-14 Cluster Range, DOT Level

	Wilks' Lambda	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Actual Test Score	.454	76.418	8.000	508	.000	.546
Analyst	.425	85.854	8.000	508	.000	.575
Regression-Estimated	.426	85.461	8.000	508	.000	.574

Table 21

Profile Analysis "Flatness" Test: 15-34 Cluster Range, DOT Level

	Wilks' Lambda	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Actual Test Score	.508	59.675	8.000	493	.000	.492
Analyst	.486	64.441	8.000	488	.000	.514
Regression-Estimated	.438	78.172	8.000	488	.000	.562

Table 22

Profile Analysis "Flatness" Test: 35-54 Cluster Range, DOT Level

	Wilks' Lambda	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Actual Test Score	.502	58.223	8.000	469	.000	.498
Analyst	.502	57.511	8.000	463	.000	.498
Regression-Estimated	.477	63.096	8.000	461	.000	.523

Table 23

Post Hoc "Flatness" Comparisons: Actual Test Score, 3 Clusters, DOT Level

Comparison	df	F	Sig.
g vs. Mean	1	140.480	.000
v vs. Mean	1	4.551	.033
n vs. Mean	1	68.291	.000
s vs. Mean	1	20.135	.000
p vs. Mean	1	.650	.421
q vs. Mean	1	2.026	.155
k vs. Mean	1	16.960	.000
f vs. Mean	1	18.307	.000

Table 24

Profile Analysis "Parallelism" Test: 2-14 Cluster Range, DOT Level

Effect			Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Actual Test Score	GATB * 3 Clusters	Pillai's Trace	.105	3.514	16	1018	.000	.052
		Wilks' Lambda	.896	3.568	16	1016	.000	.053
		Hotelling's Trace	.114	3.622	16	1014	.000	.054
		Roy's Largest Root	.102	6.513	8	509	.000	.093
Analyst	GATB * 3 Clusters	Pillai's Trace	.158	5.474	16	1018	.000	.079
		Wilks' Lambda	.846	5.525	16	1016	.000	.080
		Hotelling's Trace	.176	5.576	16	1014	.000	.081
		Roy's Largest Root	.134	8.525	8	509	.000	.118
Regression-Estimated	GATB * 3 Clusters	Pillai's Trace	.176	6.124	16	1018	.000	.088
		Wilks' Lambda	.829	6.260	16	1016	.000	.090
		Hotelling's Trace	.202	6.396	16	1014	.000	.092
		Roy's Largest Root	.173	10.991	8	509	.000	.147

Table 25

Profile Analysis "Parallelism" Test: 15-34 Cluster Range, DOT Level

Effect			Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Actual Test Score	GATB * 18 Clusters	Pillai's Trace	.419	1.627	136	4000	.000	.052
		Wilks' Lambda	.643	1.657	136	3607	.000	.054
		Hotelling's Trace	.466	1.685	136	3930	.000	.055
		Roy's Largest Root	.198	5.811	17	500	.000	.165
Analyst	GATB * 23 Clusters	Pillai's Trace	.478	1.429	176	3960	.000	.060
		Wilks' Lambda	.604	1.453	176	3701	.000	.061
		Hotelling's Trace	.534	1.476	176	3890	.000	.063
		Roy's Largest Root	.220	4.942	22	495	.000	.180
Regression- Estimated	GATB * 23 Clusters	Pillai's Trace	.532	1.603	176	3960	.000	.066
		Wilks' Lambda	.569	1.632	176	3701	.000	.068

126

Table 25 (cont'd).

Hotelling's Trace	.601	1.662	176	3890	.000	.070
Roy's Largest Root	.239	5.381	22	495	.000	.193

.

Table 26

Profile Analysis "Parallelism" Test: 35-54 Cluster Range, DOT Level

Effect			Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Actual Test Score	GATB * 42 Clusters	Pillai's Trace	.780	1.254	328	3808	.002	.097
		Wilks' Lambda	.433	1.269	328	3706	.001	.099
		Hotelling's Trace	.901	1.284	328	3738	.001	.101
		Roy's Largest Root	.285	3.313	41	476	.000	.222
Analyst	GATB * 48 Clusters	Pillai's Trace	.868	1.218	376	3760	.004	.109
		Wilks' Lambda	.392	1.230	376	3674	.003	.110
		Hotelling's Trace	1.013	1.242	376	3690	.002	.112
		Roy's Largest Root	.291	2.913	47	470	.000	.226
Regression-Estimated	GATB * 50 Clusters	Pillai's Trace	1.000	1.365	392	3744	.000	.125
		Wilks' Lambda	.336	1.381	392	3662	.000	.127
		Hotelling's Trace	1.192	1.397	392	3674	.000	.130
		Roy's Largest Root	.352	3.360	49	468	.000	.260

Table 27

Criterion Related Validity Coefficient Descriptive Statistics: SOC Level

GATB Aptitude	Mean	Std. Error of Mean	Median	Std. Deviation	Minimum	Maximum
G	.242	.0084	.250	.1371	-.14	.78
V	.181	.0087	.180	.1423	-.25	.66
N	.226	.0085	.227	.1383	-.24	.88
S	.152	.0085	.149	.1395	-.23	.55
P	.178	.0083	.167	.1361	-.25	.71
Q	.189	.0089	.187	.1448	-.37	.80
K	.148	.0081	.150	.1326	-.32	.53
F	.138	.0094	.133	.1542	-.37	.58
M	.143	.0098	.147	.1599	-.59	.57

Note. $N = 264$ for each GATB Aptitude.

Table 28

Profile Analysis "Levels" Test: 2-14 Cluster Range, SOC Level

	Source	df	F	Sig.	Partial Eta Squared
Actual Test Score	Intercept	1	910.448	.000	.777
	3 Clusters	2	5.427	.005	.040
	Error	261			
Analyst	Intercept	1	818.419	.000	.758
	3 Clusters	2	.396	.673	.003
	Error	261			
Regression-Estimated	Intercept	1	794.264	.000	.753
	4 Clusters	3	1.819	.144	.021
	Error	260			

Table 29

Profile Analysis "Levels" Test: 15-34 Cluster Range, SOC Level

	Source	df	F	Sig.	Partial Eta Squared
Actual Test Score	Intercept	1	687.015	.000	.743
	26 Clusters	25	2.034	.003	.176
	Error	238			
Analyst	Intercept	1	576.752	.000	.704
	21 Clusters	20	1.162	.289	.087
	Error	243			
Regression-Estimated	Intercept	1	680.138	.000	.738
	22 Clusters	21	1.531	.068	.117
	Error	242			

Table 30

Profile Analysis "Levels" Test: 35-54 Cluster Range, SOC Level

	Source	df	F	Sig.	Partial Eta Squared
Actual Test Score	Intercept	1	582.969	.000	.722
	39 Clusters	38	2.831	.000	.323
	Error	225			
Analyst	Intercept	1	597.272	.000	.727
	40 Clusters	39	1.139	.277	.165
	Error	224			
Regression-Estimated	Intercept	1	559.391	.000	.716
	42 Clusters	41	1.183	.221	.179
	Error	222			

Table 31

Profile Analysis "Flatness" Test: 2-14 Cluster Range, SOC Level

	Wilks' Lambda	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Actual Test Score	.346	60.020	8	254	.000	.654
Analyst	.347	59.810	8	254	.000	.653
Regression-Estimated	.373	53.110	8	253	.000	.627

Table 32

Profile Analysis "Flatness" Test: 15-34 Cluster Range, SOC Level

	Wilks' Lambda	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Actual Test Score	.359	51.596	8	231	.000	.641
Analyst	.402	43.862	8	236	.000	.598
Regression-Estimated	.368	50.351	8	235	.000	.632

Table 33

Profile Analysis "Flatness" Test: 35-54 Cluster Range, SOC Level

	Wilks' Lambda	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Actual Test Score	.415	38.411	8	218	.000	.585
Analyst	.388	42.799	8	217	.000	.612
Regression-Estimated	.374	44.914	8	215	.000	.626

Table 34

Profile Analysis "Parallelism" Test: 2-14 Cluster Range, SOC Level

Effect			Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Actual Test Score	GATB * 3 Clusters	Pillai's Trace	.148	2.543	16	510	.001	.074
		Wilks' Lambda	.855	2.580	16	508	.001	.075
		Hotelling's Trace	.166	2.617	16	506	.001	.076
		Roy's Largest Root	.140	4.451	8	255	.000	.123
Analyst	GATB * 3 Clusters	Pillai's Trace	.153	2.637	16	510	.001	.076
		Wilks' Lambda	.853	2.637	16	508	.001	.077
		Hotelling's Trace	.167	2.637	16	506	.001	.077
		Roy's Largest Root	.110	3.498	8	255	.001	.099
Regression-Estimated	GATB * 4 Clusters	Pillai's Trace	.213	2.436	24	765	.000	.071
		Wilks' Lambda	.799	2.460	24	734	.000	.072

Table 34 (cont'd).

Hotelling's Trace	.237	2.480	24	755	.000	.073
Roy's Largest Root	.140	4.455	8	255	.000	.123

Table 35

Profile Analysis "Parallelism" Test: 15-34 Cluster Range, SOC Level

Effect			Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Actual Test Score	GATB * 26 Clusters	Pillai's Trace	1.017	1.386	200	1904	.001	.127
		Wilks' Lambda	.329	1.395	200	1782	.000	.130
		Hotelling's Trace	1.223	1.402	200	1834	.000	.133
		Roy's Largest Root	.327	3.113	25	238	.000	.246
Analyst	GATB * 21 Clusters	Pillai's Trace	.862	1.467	160	1944	.000	.108
		Wilks' Lambda	.394	1.476	160	1777	.000	.110
		Hotelling's Trace	1.012	1.481	160	1874	.000	.112
		Roy's Largest Root	.274	3.324	20	243	.000	.215
Regression-Estimated	GATB * 22 Clusters	Pillai's Trace	.960	1.572	168	1936	.000	.120
		Wilks' Lambda	.350	1.591	168	1780	.000	.123
		Hotelling's Trace	1.157	1.606	168	1866	.000	.126
		Roy's Largest Root	.335	3.861	21	242	.000	.251

Table 36

Profile Analysis "Parallelism" Test: 35-54 Cluster Range, SOC Level

	Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Actual Test Score	GATB * 39 Clusters	Pillai's Trace	1.469	1.332	304	1800	.000	.184
		Wilks' Lambda	.190	1.338	304	1727	.000	.187
		Hotelling's Trace	1.887	1.343	304	1730	.000	.191
		Roy's Largest Root	.437	2.590	38	225	.000	.304
Analyst	GATB * 40 Clusters	Pillai's Trace	1.560	1.391	312	1792	.000	.195
		Wilks' Lambda	.169	1.403	312	1721	.000	.199
		Hotelling's Trace	2.047	1.412	312	1722	.000	.204
		Roy's Largest Root	.458	2.629	39	224	.000	.314
Regression-Estimated	GATB * 42 Clusters	Pillai's Trace	1.578	1.330	328	1776	.000	.197
		Wilks' Lambda	.166	1.337	328	1708	.000	.201
		Hotelling's Trace	2.063	1.342	328	1706	.000	.205
		Roy's Largest Root	.463	2.505	41	222	.000	.316

Table 37

Overall Pay Rate Descriptive Statistics

N	260
Mean	32770.50
Std. Error of Mean	924.12
Median	28500.00
SD	14900.95
Minimum	13330
Maximum	114170

Table 38

Pay Rate Descriptive Statistics: Actual Test Score Data, 3 Clusters

Cluster	N	Mean	SD	Std. Error of Mean
1	79	24475.95	7442.19	837.31
2	81	28847.16	7899.43	877.71
3	100	42501.10	17991.39	1799.14
Total	260	32770.50	14900.95	924.12

Table 39

Pay Rate Descriptive Statistics: Actual Test Score data, 26 Clusters

Cluster	N	Mean	SD	Std. Error of Mean
1	4	18550.00	6513.70	3256.85
2	25	21128.80	4483.83	896.77
3	4	37175.00	10647.84	5323.92
4	14	19735.00	3590.81	959.69
5	12	27062.50	6330.74	1827.53
6	20	29072.00	6800.39	1520.61
7	8	31133.75	8296.56	2933.28
8	10	23451.00	7743.55	2448.73
9	14	31587.86	7628.52	2038.81
10	25	29614.00	5124.26	1024.85
11	7	27277.14	8808.26	3329.21
12	9	29655.56	7397.42	2465.81
13	9	33995.56	8494.42	2831.47
14	8	34857.50	10987.31	3884.60
15	13	27192.31	9551.11	2649.00
16	9	22295.56	3998.02	1332.67
17	6	59805.00	27767.68	11336.11
18	10	42030.00	6808.66	2153.09
19	11	47267.27	22582.63	6808.92
20	7	58662.86	15751.05	5953.34

Table 39 (cont'd).

21	12	37281.67	10802.63	3118.45
22	6	40051.67	5261.94	2148.18
23	6	43703.33	7581.41	3095.10
24	6	54056.67	10654.38	4349.63
25	3	52326.67	15096.55	8716.00
26	2	91405.00	32194.57	22765.00
Total	260	32770.50	14900.95	924.12

Table 40

Pay Rate Descriptive Statistics: Actual Test Score Data, 39 Clusters

Cluster	N	Mean	SD	Std. Error of Mean
1	1	15360.00	.	.
2	3	19613.33	7540.45	4353.48
3	13	21868.46	4896.35	1358.00
4	12	20327.50	4045.25	1167.76
5	4	37175.00	10647.84	5323.92
6	3	18993.33	2507.38	1447.63
7	12	27062.50	6330.74	1827.53
8	11	19937.27	3910.85	1179.16
9	13	30446.92	7497.63	2079.47
10	8	31133.75	8296.56	2933.28
11	6	27213.33	4766.56	1945.94
12	10	23451.00	7743.55	2448.73
13	6	29451.67	5065.27	2067.89
14	6	27356.67	2223.06	907.56
15	19	30326.84	5603.84	1285.61
16	1	22350.00	.	.
17	8	33190.00	9105.15	3219.16
18	6	29095.00	8083.55	3300.10
19	6	30275.00	6880.30	2808.87
20	9	33995.56	8494.42	2831.47

Table 40 (cont'd).

21	8	34857.50	10987.31	3884.60
22	13	27192.31	9551.11	2649.00
23	8	21540.00	3520.78	1244.78
24	1	16370.00	.	.
25	6	59805.00	27767.68	11336.11
26	3	28416.67	9853.30	5688.80
27	10	42030.00	6808.66	2153.09
28	11	47267.27	22582.63	6808.92
29	4	55442.50	19828.49	9914.24
30	3	24900.00	10501.89	6063.27
31	4	39922.50	6772.81	3386.40
32	1	28340.00	.	.
33	3	62956.67	10302.28	5948.02
34	6	43703.33	7581.41	3095.10
35	9	41408.89	7498.87	2499.62
36	6	54056.67	10654.38	4349.63
37	3	52326.67	15096.55	8716.00
38	2	40310.00	791.96	560.00
39	2	91405.00	32194.57	22765.00
Total	260	32770.50	14900.95	924.12

Table 41

Pay Rate Descriptive Statistics: Analyst Data, 3 Clusters

Cluster	N	Mean	SD	Std. Error of Mean
1	64	42553.13	17600.61	2200.08
2	84	34954.88	13879.11	1514.33
3	112	25542.14	9312.37	879.94
Total	260	32770.50	14900.95	924.12

Table 42

Pay Rate Descriptive Statistics: Analyst Data, 21 Clusters

Cluster	N	Mean	SD	Std. Error of Mean
1	12	46444.17	23050.07	6653.98
2	14	47084.29	14126.03	3775.34
3	10	51874.00	15388.34	4866.22
4	7	42462.86	8418.07	3181.73
5	5	53902.00	10282.99	4598.69
6	3	63526.67	14058.12	8116.46
7	5	40842.00	14062.03	6288.73
8	6	56935.00	30049.51	12267.66
9	21	33563.33	7676.79	1675.21
10	14	33198.57	10011.26	2675.62
11	16	26337.50	11282.28	2820.57
12	13	23765.38	4566.91	1266.63
13	5	34806.00	21265.97	9510.43
14	32	29974.06	7737.72	1367.85
15	16	29811.88	11509.50	2877.38
16	7	20997.14	6043.19	2284.11
17	11	21042.73	6637.01	2001.13
18	17	31518.24	6595.76	1599.71
19	28	24044.29	5455.54	1031.00
20	11	19156.36	3565.13	1074.93

Table 42 (cont'd).

21	7	24794.29	6073.37	2295.52
Total	260	32770.50	14900.95	924.12

Table 43

Pay Rate Descriptive Statistics: Analyst Data, 40 Clusters

Cluster	N	Mean	SD	Std. Error of Mean
1	8	51741.25	26770.02	9464.63
2	6	53151.67	18606.79	7596.19
3	7	53197.14	15762.42	5957.63
4	7	42462.86	8418.07	3181.73
5	4	40850.00	813.59	406.80
6	5	53902.00	10282.99	4598.69
7	1	68640.00	.	.
8	1	52510.00	.	.
9	5	40842.00	14062.03	6288.73
10	2	69035.00	14601.76	10325.00
11	3	45113.33	8401.13	4850.40
12	2	38860.00	2899.14	2050.00
13	7	33174.29	7044.48	2662.56
14	4	44217.50	12272.10	6136.05
15	4	35850.00	7176.54	3588.27
16	3	42620.00	8486.63	4899.76
17	16	26337.50	11282.28	2820.57
18	13	23765.38	4566.91	1266.63
19	3	68756.67	42042.81	24273.43
20	3	35013.33	12175.46	7029.50

Table 43 (cont'd).

21	5	34806.00	21265.97	9510.43
22	3	35960.00	16808.70	9704.51
23	7	33361.43	6943.96	2624.57
24	14	30723.57	4974.94	1329.61
25	11	33415.45	7594.89	2289.94
26	4	31642.50	12105.39	6052.70
27	7	20997.14	6043.19	2284.11
28	8	23453.75	4537.48	1604.24
29	11	21042.73	6637.01	2001.13
30	10	30612.00	7499.06	2371.41
31	8	34806.25	10303.45	3642.82
32	9	25538.89	5225.22	1741.74
33	4	17992.50	3293.25	1646.62
34	9	21801.11	6713.63	2237.88
35	11	19156.36	3565.13	1074.93
36	7	27354.29	7916.47	2992.14
37	4	36360.00	8897.62	4448.81
38	7	24794.29	6073.37	2295.52
39	7	32812.86	5322.88	2011.86
40	10	24718.00	4145.46	1310.91
Total	260	32770.50	14900.95	924.12

Table 44

Pay Rate Descriptive Statistics: Regression-Estimated Data, 4 Clusters

Cluster	N	Mean	SD	Std. Error of Mean
1	64	48859.06	17642.41	2205.30
2	96	30123.54	9756.15	995.73
3	60	27099.17	8103.40	1046.14
4	40	21888.50	5255.73	831.00
Total	260	32770.50	14900.95	924.12

Table 45

Pay Rate Descriptive Statistics: Regression-Estimated Data, 22 Clusters

Cluster	N	Mean	SD	Std. Error of Mean
1	16	44528.13	20459.33	5114.83
2	10	56498.00	14605.80	4618.76
3	14	45097.14	8205.97	2193.14
4	5	75140.00	24190.74	10818.43
5	8	43530.00	12172.99	4303.80
6	5	42874.00	9123.14	4079.99
7	6	46646.67	19251.35	7859.33
8	13	30389.23	9518.66	2640.00
9	15	39572.00	13872.02	3581.74
10	20	30582.50	7286.09	1629.22
11	19	25697.89	6846.42	1570.68
12	8	26880.00	11197.82	3959.03
13	15	19326.67	4060.27	1048.36
14	12	23931.67	8828.92	2548.69
15	15	25646.00	7082.99	1828.82
16	19	27382.63	8249.00	1892.45
17	8	19243.75	3512.59	1241.89
18	6	24811.67	3211.27	1311.00
19	3	29483.33	2063.52	1191.38
20	14	29901.43	7277.41	1944.97

Table 45 (cont'd).

21	21	28736.19	6064.82	1323.45
22	8	24296.25	6267.19	2215.79
Total	260	32770.50	14900.95	924.12

Table 46

Pay Rate Descriptive Statistics: Regression-Estimated Data, 42 Clusters

Cluster	N	Mean	SD	Std. Error of Mean
1	16	44528.13	20459.33	5114.83
2	5	52318.00	19732.44	8824.62
3	14	45097.14	8205.97	2193.14
4	5	60678.00	6851.35	3064.02
5	1	52510.00	.	.
6	4	43440.00	14429.40	7214.70
7	3	69673.33	10384.03	5995.22
8	3	37856.67	2645.04	1527.11
9	5	42874.00	9123.14	4079.99
10	6	46646.67	19251.35	7859.33
11	5	36508.00	10215.03	4568.30
12	5	43868.00	16092.18	7196.64
13	1	114170.00	.	.
14	5	47202.00	6215.06	2779.46
15	6	26131.67	8321.18	3397.11
16	1	60910.00	.	.
17	11	31389.09	8803.43	2654.34
18	7	27872.86	8366.40	3162.20
19	8	23060.00	3518.17	1243.86
20	8	26880.00	11197.82	3959.03

154

Table 46 (cont'd).

21	3	19403.33	5936.77	3427.60
22	3	18483.33	4498.83	2597.40
23	4	29547.50	7707.71	3853.86
24	11	24520.00	6763.41	2039.25
25	4	28742.50	8013.00	4006.50
26	5	27646.00	10114.47	4523.33
27	6	21428.33	4908.28	2003.80
28	7	17828.57	3616.72	1366.99
29	4	26680.00	11559.75	5779.88
30	8	19243.75	3512.59	1241.89
31	6	24811.67	3211.27	1311.00
32	5	22708.00	10198.69	4560.99
33	4	27167.50	8938.30	4469.15
34	9	31664.44	6361.85	2120.62
35	2	27865.00	3867.87	2735.00
36	3	29483.33	2063.52	1191.38
37	14	29901.43	7277.41	1944.97
38	9	31476.67	5160.49	1720.16
39	8	24296.25	6267.19	2215.79
40	5	21378.00	3356.29	1500.98
41	9	29596.67	5217.13	1739.04
42	12	26680.83	6062.63	1750.13

Table 46 (cont'd).

| Total | 260 | 32770.50 | 14900.95 | 924.12 |

Table 47

Intraclass Correlations for Pay Data (SOC Level)

Cluster Range	Data Type	Method		95% Confidence Intervals for ANOVA Estimate
		MINQUE	ANOVA	
2-14	Actual Test Score	0.350	0.363	-0.099 - 0.826
	Analyst	0.286	0.288	-0.129 - 0.706
	Regression-Estimated	0.507	0.486	0.069 - 0.905
15-34	Actual Test Score	0.657	0.568	0.397 - 0.741
	Analyst	0.533	0.451	0.255 - 0.648
	Regression-Estimated	0.591	0.520	0.339 - 0.703
35-54	Actual Test Score	0.630	0.561	0.411 - 0.712
	Analyst	0.534	0.490	0.339 - 0.643
	Regression-Estimated	0.634	0.573	0.431 - 0.716

Note. MINQUE = Minimum norm quadratic unbiased estimation method.
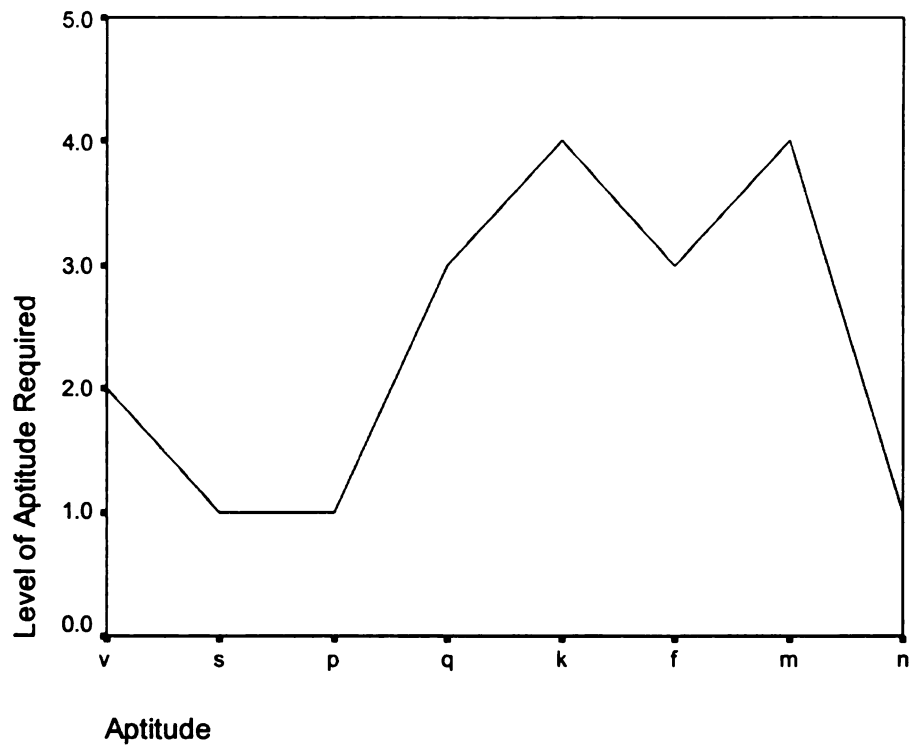
Figure 1

Example of a Profile – Occupational Aptitude Requirements for Architects (1-5 scale)

Note. v = verbal aptitude, s = spatial aptitude, p = form perception, q = clerical, k = motor coordination, f = finger dexterity, m = manual dexterity, and n = numerical aptitude.
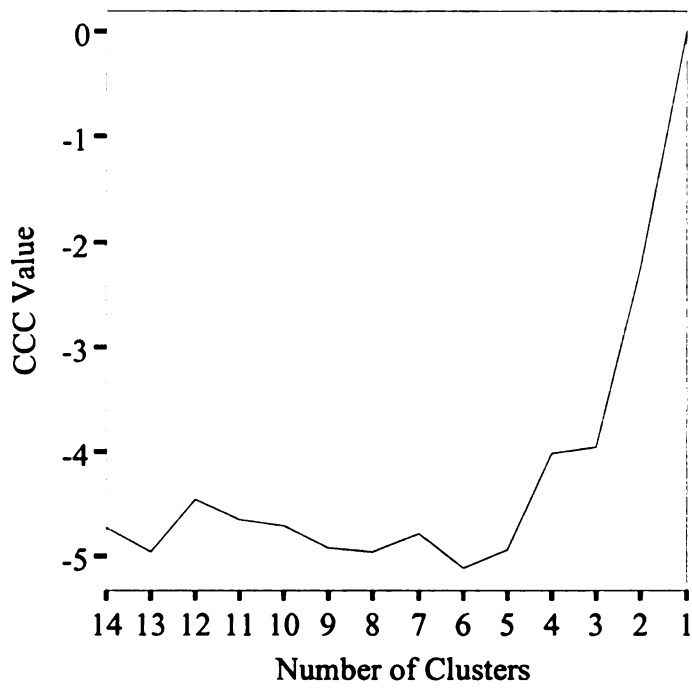
Figure 2

CCC values for actual test score data (SOC level) for 1 to 14 clusters
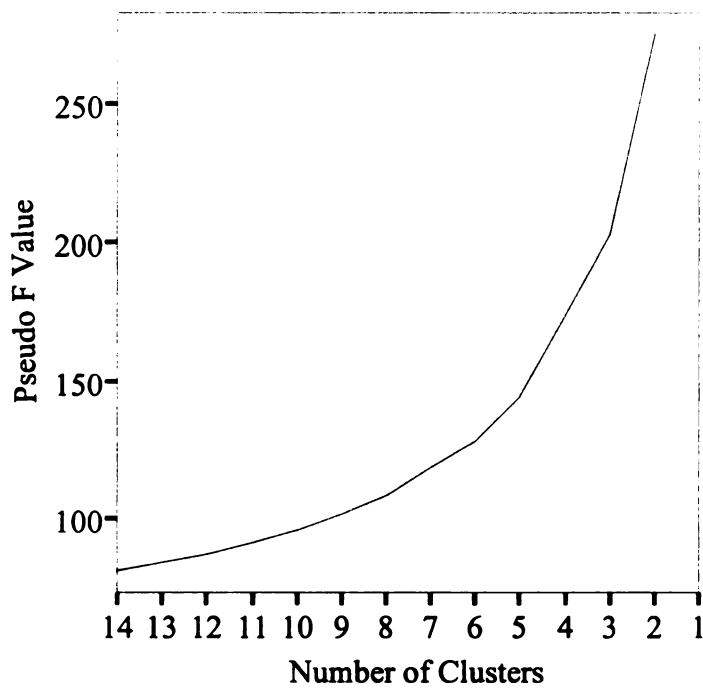
Figure 3

Pseudo F values for actual test score data (SOC level) for 1 to 14 clusters
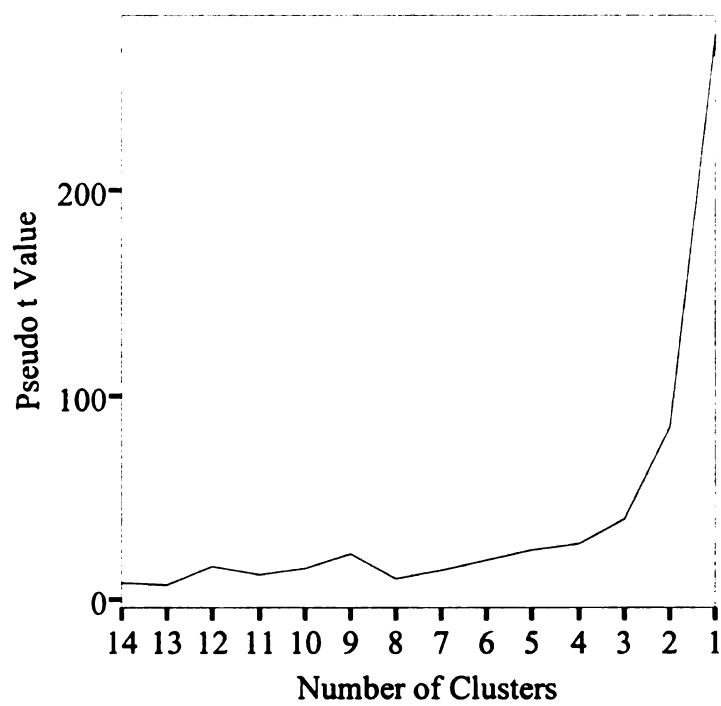
Figure 4

Pseudo t^2 values for actual test score data (SOC level) for 1 to 14 clusters
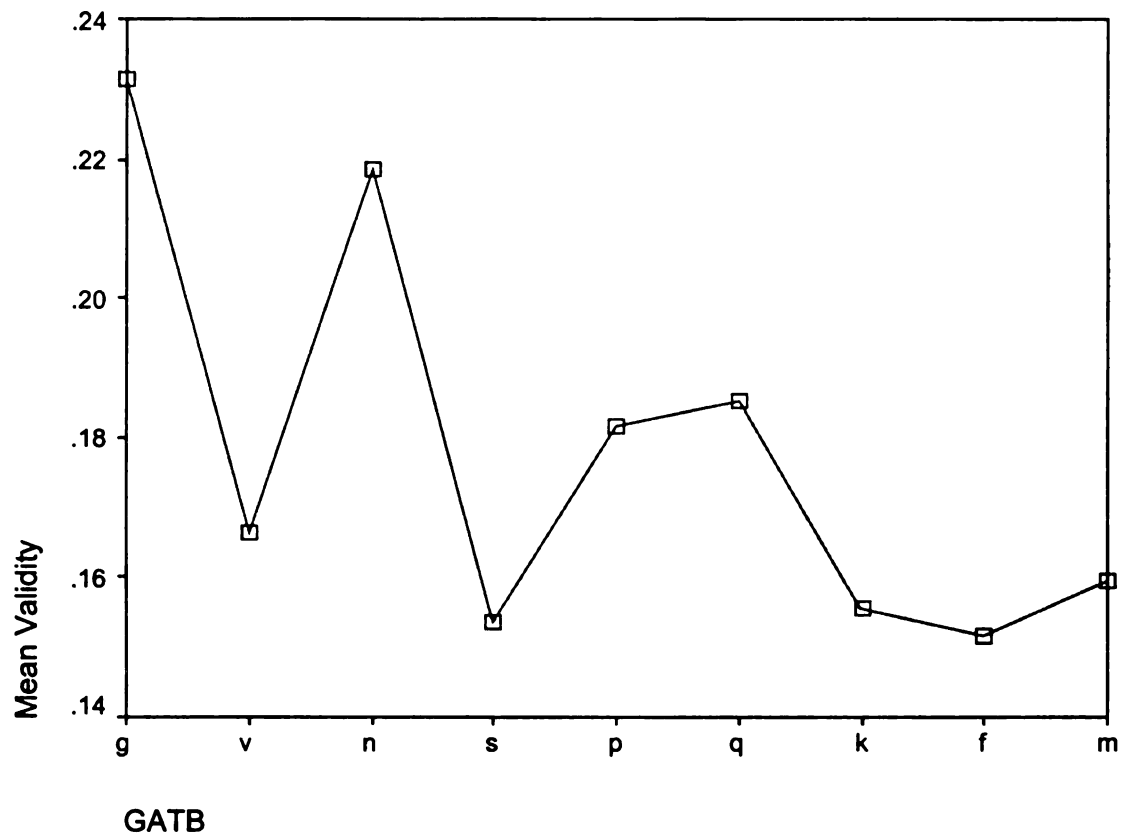
Figure 5

Mean validity profile across all clusters for actual test score data (DOT level) for the 2-

14 cluster range
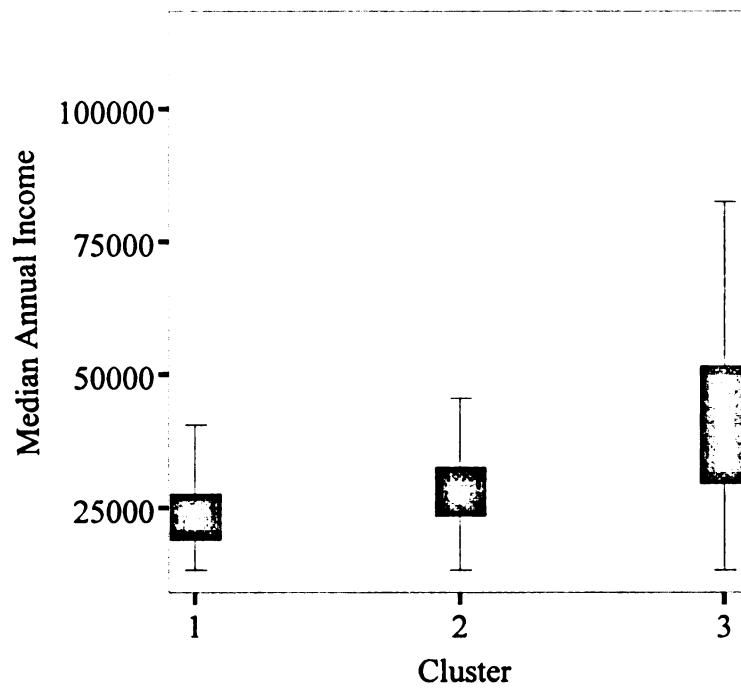
Figure 6

Pay data boxplots for actual test score data, 3 clusters
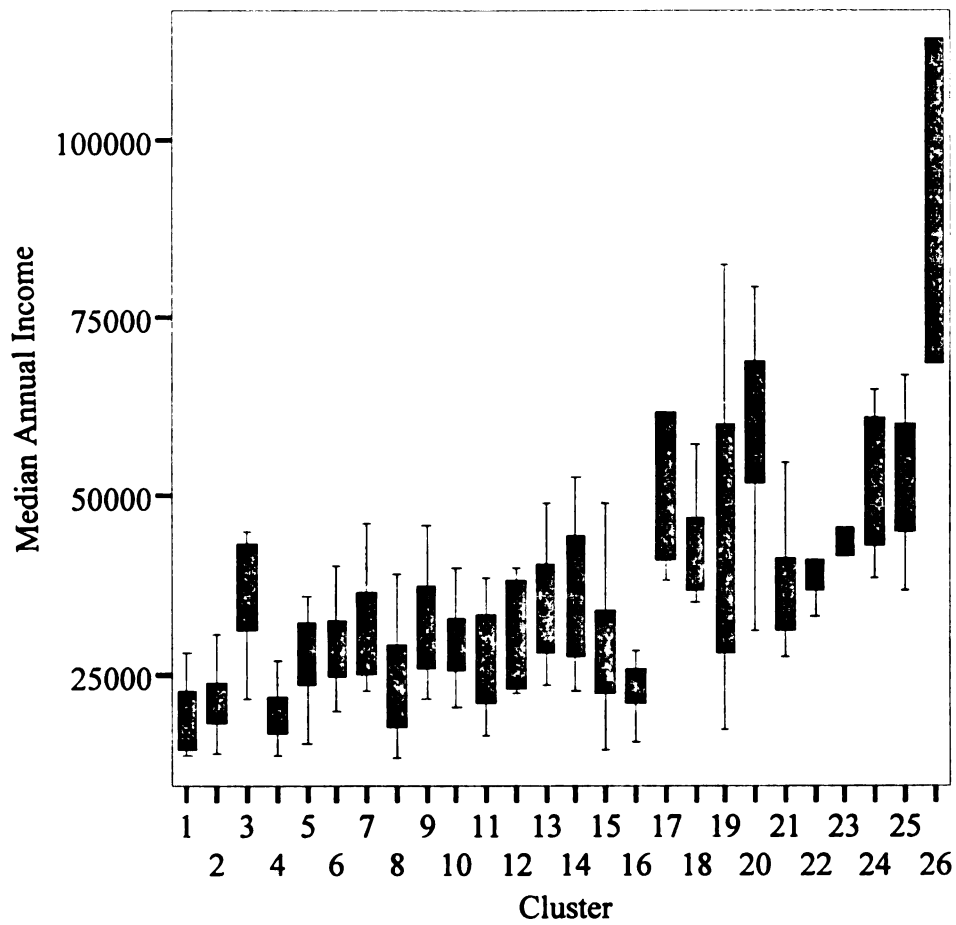
Figure 7

Pay data boxplots for actual test score data, 26 clusters
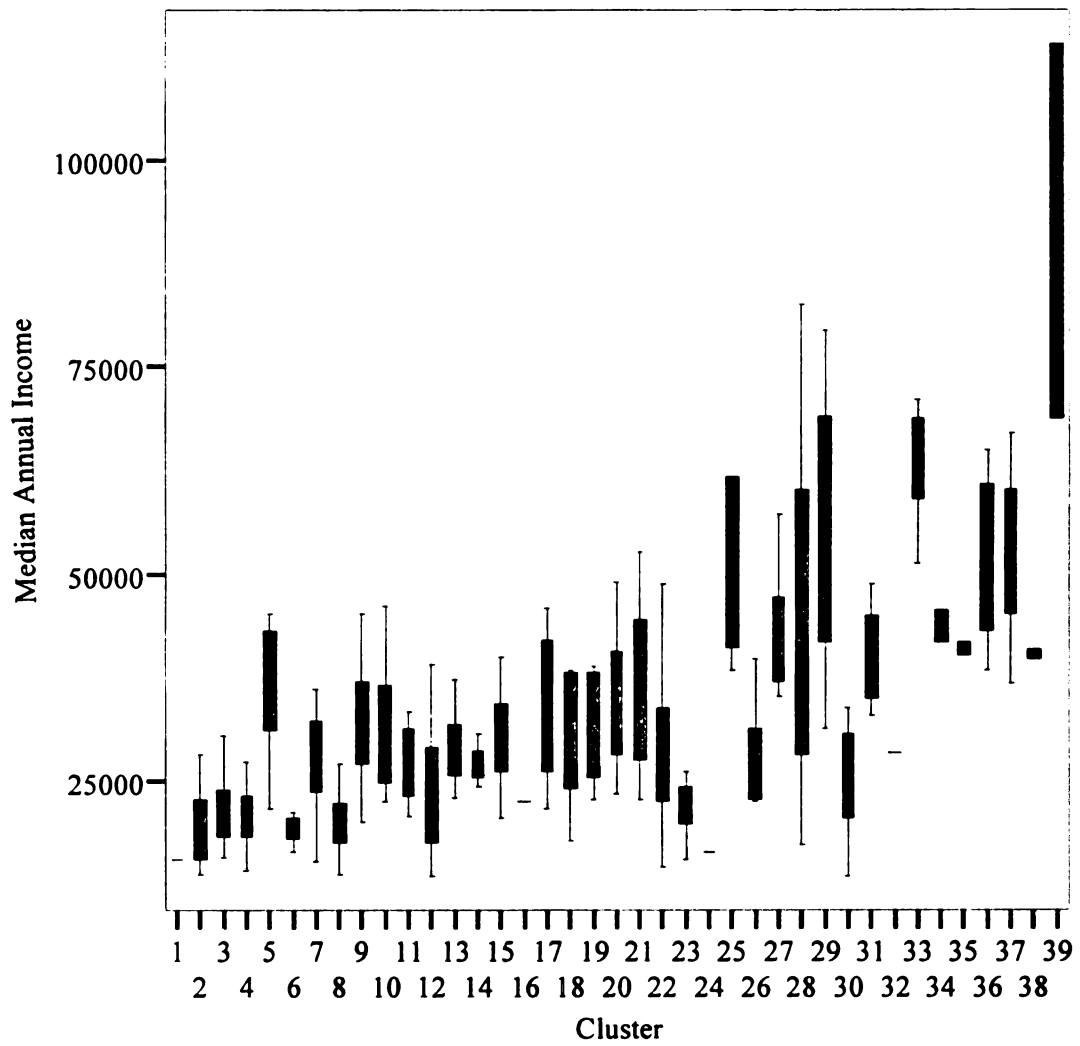
Figure 8

Pay data boxplots for actual test score data, 39 clusters
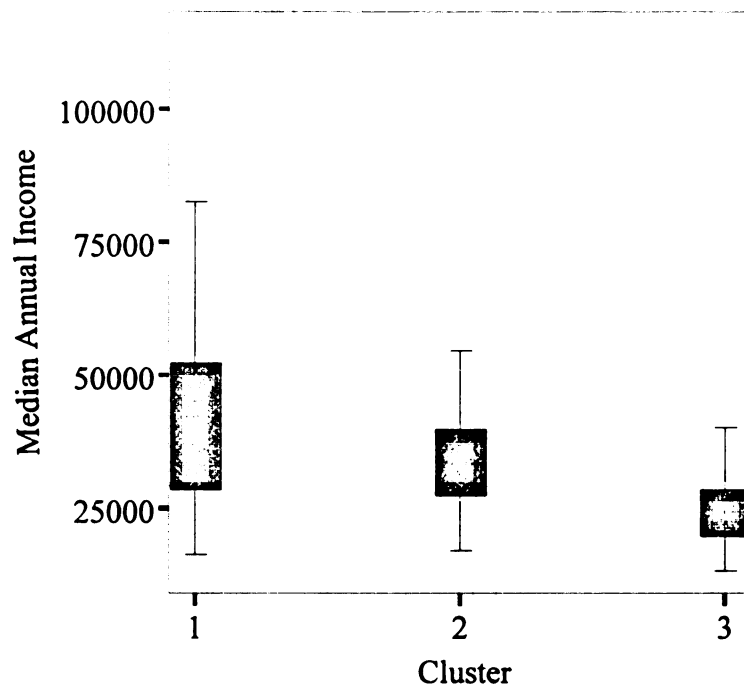
Figure 9

Pay data boxplots for analyst data, 3 clusters
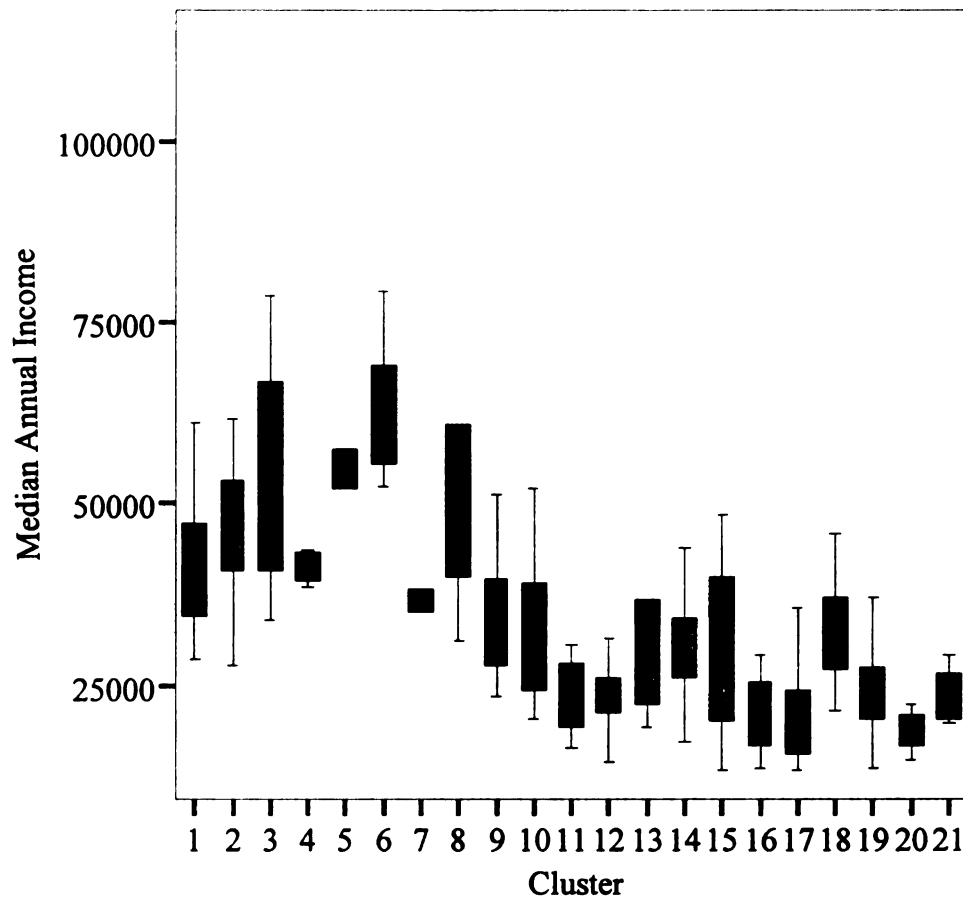
Figure 10

Pay data boxplots for analyst data, 21 clusters
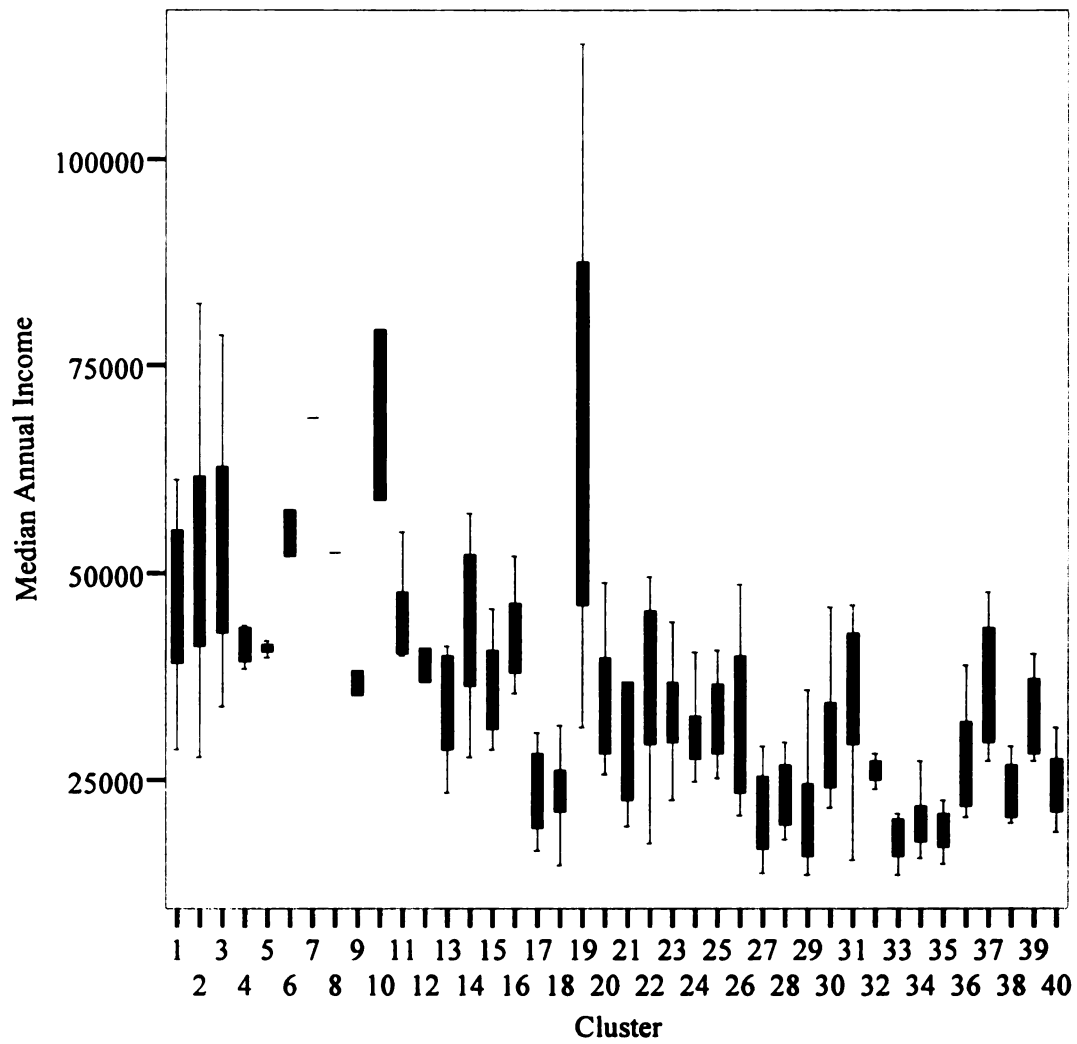
Figure 11

Pay data boxplots for analyst data, 40 clusters
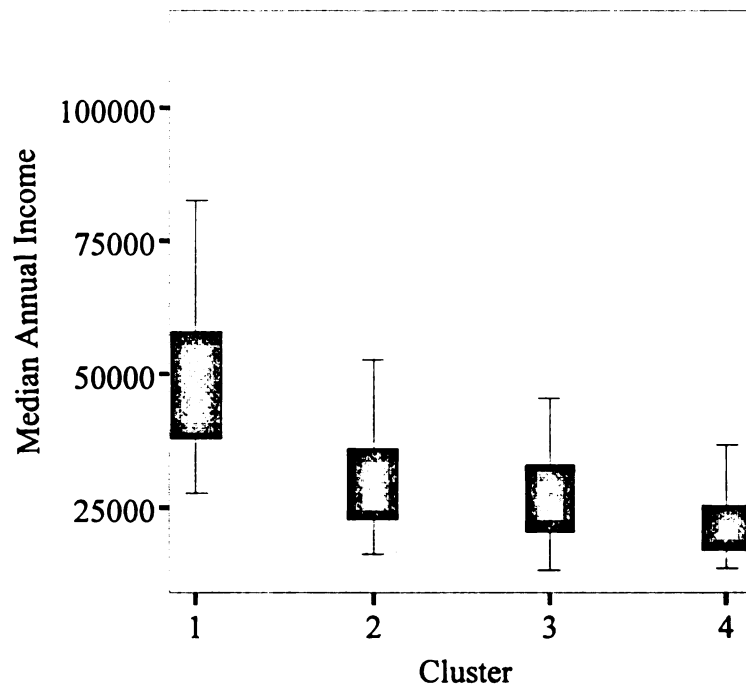
Figure 12
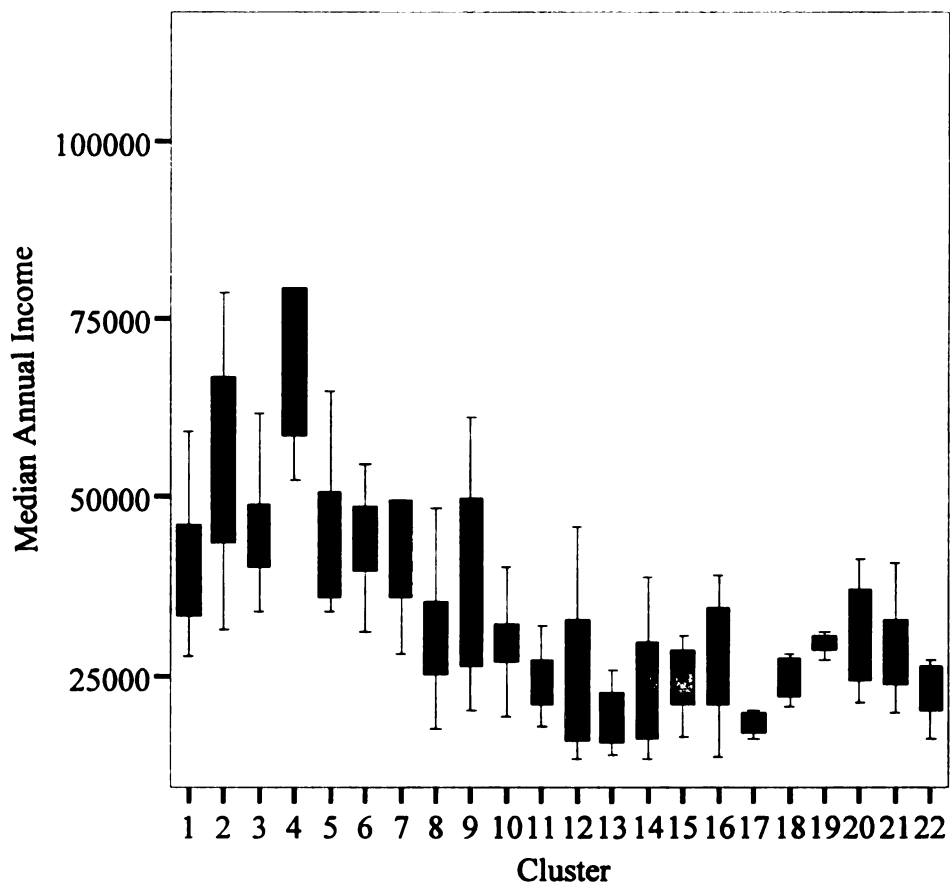
Pay data boxplots for regression-estimated data, 4 clusters

Figure 13

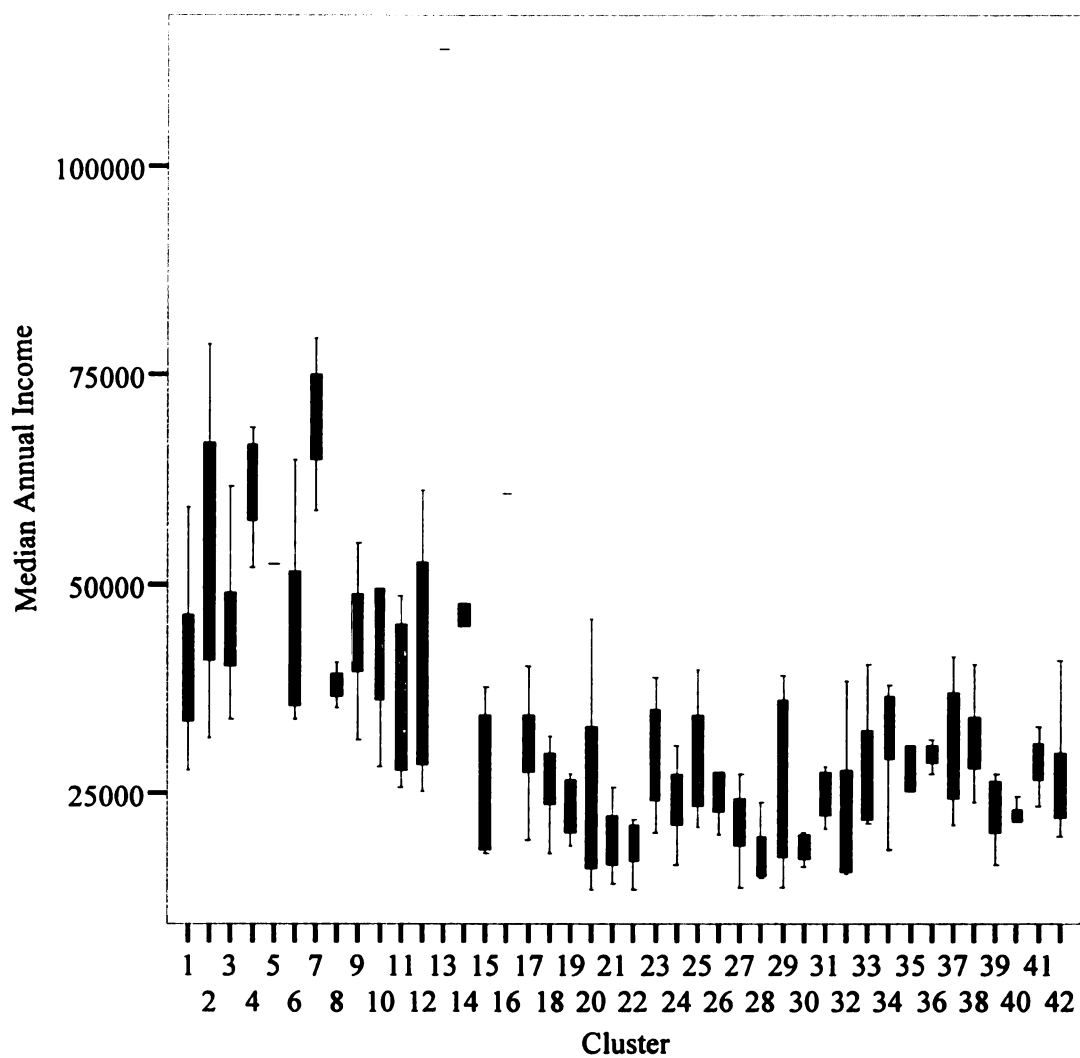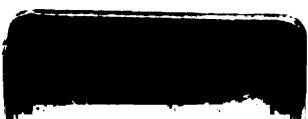*Pay data boxplots for regression-estimated data, 22 clusters*

Figure 14

Pay data boxplots for regression-estimated data, 42 clusters