Psychology Science, Volume 46, 2004 (1), p. 99-127

The effects of data type on job classification and its purposes

PATRICK D. CONVERSE[1], FREDERICK L. OSWALD

Abstract

Often, organizational psychologists use methods of classification to group individual jobs into larger groups that are then used for personnel-related functions. In classifying jobs, researchers and practitioners make several important considerations, such as deciding on the type of data collected and choosing the appropriate quantitative clustering method. We highlight some of these considerations with a dataset comprising three different types of ability data. Complementing previous research, these three ways of measuring ability requirements resulted in substantially different job clusters. Although these differences did not have clear implications for the practical purposes examined – test validation, job evaluation, and career exploration – it was evident that the differences influenced the final results obtained and therefore would influence personnel decisions made from them.

Keywords: Aptitude profiles, Job analysis, Clustering methods, Job classification, Job families

[1] Pat Converse, 129 Psychology Research Building, East Lansing, MI 48814-1117;
E-mail: convers8@msu.edu

The Effects of Data Type on Job Classification and Its Purposes

In organizational psychology, research and practice often require classifying individual positions or jobs into groups whose elements share similar characteristics. For example, individual positions may be grouped into jobs, or jobs may be grouped into larger "job families" based on similarities in job requirements, such as having similar profiles of different abilities required for adequate job performance. This classification process is similar in spirit to factor analysis, in the sense that it results in a small yet sensible number of groups to simplify and amplify relevant similarities and differences. Rather than serving as an end in itself, however, job classification is a tool that assists in other personnel-related functions (Pearlman, 1980). For example, classifying or clustering jobs into larger job families can play a critical role in activities such as appraising employees' job performance (e.g., Cornelius, Hakel, & Sackett, 1979), validating employee selection tests (Arvey & Mossholder, 1977), evaluating jobs (Pearlman, 1980), planning career paths (Harvey, 1986), and counseling individuals seeking vocational guidance (e.g., Peterson, Mumford, Borman, Jeanneret, & Fleishman, 1999).

There are many benefits of clustering jobs effectively. For instance, rather than having to develop distinct measures to assess employee performance for each individual job in an organization, job clustering can justify developing measures for a smaller number of job groups. Clustering takes what would originally be a cumbersome, costly, and time-consuming task and transforms it into a more manageable task that is less expensive and less time consuming, yet hopefully just as useful. Of course, the ability to reduce the total number of jobs to a manageable and appropriate number of job families assumes that (a) individual positions – and individual jobs – can be aggregated into job families on relevant characteristics that apply across jobs, and (b) losing unique information about individual positions and individual jobs through job grouping does not adversely affect the purposes to which the job clusters are put.

Researchers and practitioners classifying jobs end up addressing – either implicitly or explicitly – a number of important considerations that affect the resulting job clusters and, consequently, the decisions and outcomes that result from them. These include:

(a) deciding on the set of constructs (and their related measures) on which jobs will be clustered,
(b) paying attention to the characteristics of the sample and measurement methods to understand potential method variance and profile-irrelevant sample dependencies,
(c) scaling the data within jobs and/or within variables (e.g., centering or standardizing),
(d) selecting the metric used to reflect similarity between jobs, and
(e) choosing the appropriate method for clustering jobs.

The present study complements a substantial amount of research related to the advantages and disadvantages of various quantitative clustering techniques (e.g., Colihan & Burger, 1995; Milligan, 1981; Milligan & Cooper, 1987) by focusing on issues related to the constructs or types of data used for job classification.

Related research has examined the effects of different types of data on results when clustering the same set of jobs based on data from *different* psychological domains (e.g., grouping on ability requirements versus grouping on job tasks performed). The present study complements and extends this research by focusing on the potential effects on results due to clustering different types of job characteristic data that come from a *single* domain, namely abilities. This study reflects an exploratory empirical approach, 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 desired purposes and outcomes. Specifically, we have three types of measures of the same abilities: (1) mean employee ability test scores, (2) expert job analyst ratings of required ability levels, and (3) regression-estimated ability scores. We then cluster the profiles of each of these different measures of abilities for the same set of jobs. In research and practice only one type of data is usually available for a particular construct of interest. The present study has the advantage of investigating separate cluster solutions using different types of data for the ability constructs of interest, allowing us to infer the extent to which results might have differed in a job clustering study had a different type of data been available.

Clusters resulting from these three types of ability measures will be examined in terms of their effectiveness for three different purposes: selection test validation, job evaluation, and career exploration. In any practical situation, effective cluster solutions need to be internally consistent: all jobs are placed in only one cluster, each cluster contains similar jobs in terms of the relevant dimensions, and there are meaningful distinctions between different clusters (e.g., see Pearlman, 1980). Although these characteristics are desirable, they do not ensure effectiveness in light of a given purpose, and therefore relevant data *external* to the clustering process can be examined to determine the usefulness of a given cluster solution. How usefulness or effectiveness is determined depends on the particular context, but generally speaking this determination often relates to the extent to which a cluster solution assists in making accurate or appropriate decisions. For example, job clusters used in selection test validation are more useful if they allow for accurate decisions regarding the criterion-related validity of a given measure for all (or most) of the jobs in each cluster; less effective clusters are those leading to inappropriate conclusions regarding validity levels for many jobs. This will be discussed further in later sections.

Same Jobs, Different Construct Domains

Previous organizational research examining the effects of data type on job clustering suggests that data from *different* construct domains can result in substantially different job clusters (e.g., Cornelius, Carron, & Collins, 1979; Ghiselli, 1966; Pearlman, 1980). For instance, Cornelius et al. (1979) found that, in a sample of seven foreman jobs, different job clusters resulted depending on whether the data were ratings of the tasks performed on the job, Position Analysis Questionnaire (PAQ) dimensions reflecting the behaviors required to perform the job, or measured worker abilities. Although the sample of jobs here is very small and not generalizable, the major implication of these findings is that the objective of classification should, first and foremost, determine which domain of variables is relevant for developing job characteristic profiles. If cluster solutions differ depending upon the type of data used, then the type of data should be chosen carefully according to the objective of clustering.

However, a few studies report exceptions to this general finding of cluster differences (e.g., Baughman, Norris, Cooke, Peterson, & Mumford, 1999; Hartman, Mumford, & Mueller, 1992). For example, Hartman et al. (1992) 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 similar job classifications can be developed using these different job descriptors. Nonetheless, these studies do not report completely consistent classification results, suggesting there may well be important differences in the classifications produced by different data types.

Same Jobs, Same Construct Domains

Less research attention has focused on how even from within the *same* construct domain job groupings are affected by different types of profile data. Types of profile data can differ both in the specificity of the constructs measured as well as in how those constructs are measured. For example, within the abilities domain, different constructs such as verbal ability, numerical ability, or motor coordination can be measured, and each construct can be measured in different ways, such as by ability tests, from samples of employees' work behavior reflecting ability, or via expert ratings of ability requirements of the job.

Little is known about the extent to which job profile data might yield different job groupings when there are differences in how the constructs in the profile are measured. Garwood, Anderson, and Greengart (1991) explored a similar issue by examining how the degree of task overlap across jobs, the number of tasks performed in jobs, and the number of people in each job affected different types of quantitative clustering methods. Their results, based on simulated job analysis data, indicated that clustering method effectiveness varied as a function of the properties of these data. Although this study examined the general issue of how clustering results are materially affected by the type of data used within a given construct domain, these data resulted from different job analysis situations (e.g., situations where there is no task overlap across jobs versus situations where there is substantial task overlap), rather than from different measurement methods. The present study therefore complements and extends Garwood et al. (1991) by (a) examining the effects of different types of data resulting from different measurement *methods* on job clustering and (b) examining the implications of these effects for a few major *purposes* to which clusters are put in organizational research and practice.

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 issue. O*NET is the U.S. 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 psychological and situational characteristics about work and the worker (Dye & Silver, 1999). Among the resources provided by O*NET (see www.onetcenter.org), the O*NET Ability Profiler helps individuals just entering careers or in career transition focus their career-search activities. The Ability Profiler uses subtests of the well established General Aptitude Test Battery (GATB; U.S. Department of Labor, 1979) to measure individuals' 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 uses statistical formulas to compare individuals' ability profiles with the profiles of ability requirements for jobs, presenting each individual with a subset of jobs that most closely fits his/her profile.

In its development, the O*NET's Ability Profiler used three types of ability data: actual GATB test score profiles, job analyst-rated GATB profiles, and regression-estimated GATB profiles, all of which are described in more detail below. Again, the present study examines how profiles of these different types of ability data affect job groupings. That is, although all three types of ability data represent measurements of the same nine GATB aptitudes, the data come from very different sources, and different processes give rise to these data. These substantive differences in these data may result in different ability profiles among the three types of data for the same jobs, which in turn may lead to different job groupings. Differences in job groupings may then have practical implications for the many purposes to which the groupings might be put.

Three Types of Ability Data

Actual GATB test score profiles. Actual GATB profiles were obtained from the test scores of workers with each of the nine 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 employee test scores are assumed to reflect the ability levels actually required for satisfactory performance based on evidence indicating that over time individuals tend to select themselves into jobs commensurate with their ability levels (e.g., Wilk, Desmarais, & Sackett, 1995; Wilk & Sackett, 1996). Actual GATB ability profiles exist for 545 jobs where workers were tested with the GATB.

Job analyst-rated profiles. Job analyst profiles of ability requirements come from the *Dictionary of Occupational Titles* (*DOT*). These rating data were gathered before the development of the O*NET. 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). We used job analyst ratings from the 1991 revised fourth edition (the most recent edition of the *DOT*), although job analyst rating data are currently being collected for the O*NET database. To develop job aptitude profiles, expert job analysts first observed individual jobs and described the job's tasks and purposes. On the basis of these descriptions and other observations, analysts then rated each job on 11 aptitudes: the nine previously mentioned 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 level of the 12,000+ jobs described in the *DOT* such that each *DOT* job's rating on each of the 11 aptitudes reflects the modal value of the ratings from its constituent jobs (Cain & Green, 1983).

Regression-estimated profiles. Although actual GATB test score profiles existed for 545 *DOT*-level jobs, these jobs did not cover the entire population of jobs in the O*NET. Therefore, McCloy et al. (1999) used the actual GATB profiles, as well as data from the *DOT*, to develop regression-estimated ability profiles for all 12,000+ *DOT* jobs. This required three steps. First, 48 predictor variables, constituting *DOT* job analysis information such as job analysts' ratings of how much a given job deals with Data, People, and Things (see Table 1), were reduced to seven promax-rotated component scores. Second, actual GATB mean test scores, for the 545 jobs providing them, were regressed on these component scores (regressing onto the original 48 variables would have capitalized on chance, leading to the regression model overfitting the data). Resulting regression weights were then used to develop predicted GATB scores for each *DOT*-level job – including those jobs without actual GATB test scores.

Table 1:
DOT Variables Used to Predict GATB Scores

Data, People, Things	Temperaments
	• Directing
Reasoning, Math, Language	• Repetitive
	• Influencing
Specific Vocational Preparation	• Variety
	• Expressing
Physical Demands	• Stress
• Strength	• Tolerances
• Climbing	• Under
• Balance	• People
• Stooping	• Judgments
• Kneeling	
• Crouching	GATB Aptitude Ratings
• Crawling	• G – General Intelligence
• Reaching	• V – Verbal Ability
• Handling	• N – Numerical Ability
• Fingering	• S – Spatial Ability
• Feeling	• P – Form Perception
• Talking	• Q – Clerical Ability
• Hearing	• K – Motor Coordination
• Tasting/Smelling	• F – Finger Dexterity
• Near Acuity	• M – Manual Dexterity
• Far Acuity	• E – Eye-Hand-Foot Coordination
• Depth Perception	• C – Color Discrimination
• Accommodation	
• Color Vision	
• Field of Vision	

To summarize, the O*NET Ability Profiler's development involved three distinct types of ability profiles: actual test score profiles, expert job analyst-based profiles, and regression-estimated profiles. Although the three types of profiles are intended to measure the same aptitudes, the different substantive processes that led to each type of profile may result in different profiles among the three types of data, even for the same job. Test-based data come from the responses of multiple employees who took an aptitude test battery, and within each job their scores were averaged to yield job-level estimates of aptitude requirements. Job analyst-based data reflect a process of cognitive estimation in which job analysts were required to observe, encode, store, retrieve, and integrate job relevant information. Regression-based data come from job analysis information predicting required abilities. The regression data reflect the characteristics of rating data (and their measurement errors) along with the assumptions of the regression model (and their prediction errors). Each type of data thus has unique strengths and weaknesses in terms of providing accurate estimates of ability requirements for each job. For example, analyst ratings may have been influenced by the information processing limitations and biases that can be present in any subjective judgment task (e.g., order and contrast effects; see Morgeson & Campion, 1997), whereas actual test scores may have been influenced by other biases related to test-taking situations (e.g., test-taker motivation), and regression-estimated scores may have been influenced by still other biases (e.g., the types of predictor variables used).

Implications for Personnel-Related Functions

The brief discussion of the three data types suggests that ability estimates that result from these different methods may well differ. Consequently, these three types of data may not result in the same (or even similar) profiles or job clusters, which in turn has practical implications given the particular purpose for job clustering, such as test validation or job evaluation. Thus, creating job clusters requires not only choosing the broad psychological domain and quantitative method with care, according to the objective of classification; it also requires choosing, or at least being mindful of, the type of data within a given domain so that the cluster analysis can yield useful job clusters.

Job clustering serves a wide variety of important purposes in organizations, as Table 2 shows. However, job clusters based on ability requirements are only appropriate for some of these purposes. Specifically, ability-based job clusters may tend to be appropriate when jobs are classified for test validation, vocational and educational guidance, job placement, personnel classification, internal job classification, and job evaluation. Ability-based job families might be useful in vocational guidance situations, for instance, 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 employee selection tests. The present study evaluates relative strengths and weaknesses of test-based, analyst-based, and regression-based ability data by examining the effectiveness of clustering solutions resulting from these data for three major purposes for which criteria are available: test validation (personnel selection), job evaluation, and career exploration with O*NET's Ability Profiler.

Table 2:
Objectives of Clustering Jobs

- Test validation (personnel selection; Arvey & Mossholder, 1977)
- Job evaluation (for setting pay structures, wage and salary administration; Pearlman, 1980)
- Vocational and educational guidance (Pearlman, 1980)
- Job placement (Pearlman, 1980)
- Personnel classification (Pearlman, 1980)
- Establishing career promotion ladders (career-path planning) and lines of job transfer (Pearlman, 1980)
- Internal job classification (Pearlman, 1980)
- Exploratory research, theory development, and methodological research objectives (Pearlman, 1980)
- Performance appraisal (e.g., Cornelius, Hakel, & Sackett, 1979)
- Establishing vocational training curricula (Pearlman, 1980)
- Developing training programs (Pearlman, 1980)
- Population-level occupational data collection and analysis for economic and social purposes (Pearlman, 1980)

Test validation. Ability-based job clusters can be useful when seeking to establish the criterion-related validity of an employment test (Arvey & Mossholder, 1977). For example, clustering several jobs with similar ability requirements may be required to yield a large enough sample for statistically powerful validation. Even when sample size is not a concern, combining jobs with similar ability requirements may still be desirable. For instance, instead of developing and validating several distinct selection tests for superficially different jobs, organizations can validate a smaller number of ability tests for job clusters, thereby simplifying the test validation process.

Comparing the criterion-related validity coefficients associated with the jobs in each cluster is one way to examine the relative strengths and weaknesses of clusters resulting from test-based, analyst-based, and regression-based data for use in test validation. Specifically, for test validation purposes, it would be desirable to have job clusters consisting of jobs with relatively homogeneous criterion-related validities. Conversely, if validity coefficients were heterogeneous, then job clusters would mask important between-job differences in predictor-criterion relationships, and one might reach the undesirable conclusion that a predictor is valid (or not valid) for all the jobs in a cluster, when in fact the magnitudes of the validity coefficients differ across jobs. For this reason, more useful clusters in this context are not necessarily those with the highest mean criterion-related validity coefficients, but rather those with meaningful between-cluster yet little within-cluster variability in validity coefficients. Therefore, we compare the utility of test-based, analyst-based, and regression-based job clusters for test validation by examining the amount of between-cluster variability relative to within-cluster variability in criterion-related validity coefficients across cluster solutions. Higher within-cluster variability indicates more heterogeneity and relatively less utility.

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 appropriate and useful for determining occupational pay levels. For instance, it may be appropriate for an organization to determine salary, in part, based on ability requirement levels such that individuals in jobs requiring higher ability levels tend to receive higher pay (Milkovich & Newman, 1990). Therefore, jobs might be clustered according to ability requirements, and the jobs within each cluster are paid similarly because they require similar levels of abilities. In this case, sensible and useful ability-based job clusters are those with little within-cluster variability in pay rates. The usefulness of job clusters based on test-based, analyst-based, and regression-based data for job evaluation can therefore be compared by examining variability in pay levels across cluster solutions, where 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 have implications for career exploration with O*NET's Ability Profiler. If job clusters obtained from the actual GATB profiles are substantially different from those obtained from the regression-estimated profiles, this indicates that the types of jobs that the Ability Profiler encourages clients to pursue may be somewhat data-dependent. 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 were based on a different type of ability data, meaning that the type of ability data used is important and not interchangeable.

Summary

The present analysis will reveal differences and similarities in job clusters using three different types of ability data: employee test score data, job analyst data, and regression-estimated data. Analyses will examine the within- versus between-cluster variability in profiles of criterion-related validity coefficients and in pay rates, for test validation and for job evaluation purposes respectively. Large between-cluster and little within-cluster variability indicates relatively useful job clusters. Note that this criterion does not necessarily determine the effectiveness of a particular clustering method a priori, as it is not clear which method would produce these outcomes given that ability profiles are used to cluster jobs rather than the criteria themselves. For instance, a clustering method that minimizes within-cluster variability in abilities will not necessarily minimize within-cluster variability in pay rates. Additionally, similarities and differences in clusters resulting from these three data types are examined to reveal potential implications for career exploration with O*NET's relatively new Ability Profiler. Clearly there are no tidy prescriptions for profile matching, but past research informs an appropriate analytic approach that was taken in the present study. Other types of analyses would certainly be possible, but our particular approach is detailed below.

Method

Job clustering requires not only deciding how to cluster jobs, but also how to determine the number of job clusters supported in the dataset. Research on quantitative clustering methods indicates that no single technique for grouping jobs or determining the number of clusters present is superior in all situations, although a few tend to perform well under many circumstances (e.g., see Colihan & Burger, 1995; Harvey, 1986; Milligan & Cooper, 1987; Milligan, 1981; Milligan & Cooper, 1985). In particular, in terms of how individuals (or jobs) are grouped, Ward's (1963) minimum variance technique is recommended (e.g., Milligan & Cooper, 1987). The present study uses Ward's method in forming clusters by minimizing the total within-group or within-cluster sum of squares (i.e., the sum of the squared deviations of the scores about their mean). Alternatively, Ward's method can be employed using correlations, where minimizing within-cluster sum of squares is replaced by minimizing $1-R^2$.

Similarly, in determining the appropriate number of clusters present in the data, there are reasonable alternatives but no prescriptions. Research by Milligan and Cooper (1985) suggests three criteria: the cubic clustering criterion (CCC; see Sarle, 1983), the variance ratio criterion or pseudo F statistic (Calinski & Harabasz, 1974), and the pseudo t^2 statistic (see Duda & Hart, 1973). Each index of adequacy for the number of clusters incorporates several types of statistics for each step in the clustering process using Ward's method. Generally speaking, the indices reflect whether two clusters joined at a given step in the clustering process should in fact be combined. Examining these values as the number of clusters in the clustering process gets smaller helps one decide on the number of clusters present in a given dataset: There should not be too few clusters that the data are not well represented, but there should not be so many clusters that the cluster solution is too complex and overfits the data. Others have suggested that convergence across these three statistics helps determine the number of clusters present (e.g., SAS Institute, 1999). Specifically, local peaks in the CCC and pseudo F statistic, along with a small value for the pseudo t^2 followed by a larger t^2 at the next clustering step, suggest that the appropriate number of clusters has been identified.

Data

All three types of ability data describe jobs at the level of the *Dictionary of Occupational Titles* (*DOT*). Currently, however, all U.S. government agencies collecting occupational information are moving over to the Standard Occupational Classification (SOC) system. The *DOT* and SOC systems of organizing jobs differ in several ways, but in the present context a primary difference is that the SOC is a broader classification system containing approximately 820 classifications that subsumes the *DOT* system, which contains over 12,000 job classifications. Given that the SOC system will be used by U.S. government agencies for all future job-related data collection, we aggregated *DOT* data up to the SOC level. The SOC contains hundreds of classifications and is likely to be refined enough to warrant useful job classification, but in exercising caution, we clustered and analyzed jobs at both levels.

Missing data reduced the working data set from 545 to 518 *DOT*-level jobs. We did not impute missing data because (a) 518 jobs constituted 95% of the jobs as well as a sufficiently large sample for our purposes, and (b) our focus in this article is on comparing different types of data in job clustering, and therefore we felt it was more appropriate to include only the raw data, as inclusion of missing-data estimates could lead to less accurate results and conclusions. In addition, aptitude G (General Intelligence) was excluded from analyses because it is redundant with GATB aptitudes V, N, and S. Analyses were conducted on test-score, job analyst-based, and regression-estimated ability profiles consisting of the eight remaining GATB aptitudes: V, N, S, P, Q, K, F, and M (see Table 1).

To conduct SOC-level analyses, the 518 *DOT*-level jobs were placed into their 264 corresponding SOC categories. We then averaged *DOT*-level ability scores within each SOC classification, yielding test-score, job analyst-based, and regression-estimated ability profiles, with each type of profile comprising the same eight aptitudes. Profiles were then analyzed separately at both the *DOT* and SOC level.

Table 3:
Number of SOCs within Each Major Group

Major Group	Description	*f*
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
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

Note. f = Number of SOCs in the Major Group.

To facilitate classification, the SOC system divides jobs with similar skills and work activities into 23 Major Groups, 96 Minor Groups, and 449 Broad Occupations (Bureau of Labor Statistics, 2001a). Table 3 shows that the SOCs covered by *DOT* jobs in the present dataset cover 22 of the 23 Major Groups. Furthermore, given any particular SOC job, the percentage of *DOT*s with data ranges from 0.4% (1 *DOT*-level job out of 251) to 100% (1 *DOT*-level job with data where in fact only 1 *DOT* job fits into the SOC), with a mean of 30% and a (large) standard deviation of 29%. Thus, the extent to which *DOT*s with data represent all *DOT*s within each SOC varies considerably across SOCs. These percentages should not be strictly interpreted, however, because *DOT* job titles themselves vary in their breadth. However, the results do indicate that *DOT* data are well represented and distributed across the SOC system.

Results

Descriptive Statistics and Reliability

Tables 4, 5, and 6 provide descriptive statistics and intercorrelations for actual test score, job analyst, and regression-estimated GATB data at the *DOT* level. Tables 7, 8, and 9 provide similar results at the SOC level.

Previous studies indicate that the GATB aptitudes are measured reliably across numerous populations and contexts. For example, studies from high school, college, and adult samples using test-retest intervals of one day to three years generally produced reliability coefficients in the range of .80 to .90 (U.S. Department of Labor, 1970). Thus, the GATB test score data in the present study should be highly reliable, especially because profiles reflect mean test scores across employees within a job, and means are more stable than individual scores.

Unfortunately, less is known about the reliability of job analyst 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 [*DOT*] ratings during the course of fourth edition production” (p. 169). However, a few researchers have estimated the reliability of *DOT* ratings using ratings based on procedures very similar to those used to generate actual *DOT* ratings (e.g., Cain & Green, 1983; Geyer, Hice, Hawk, Boese, & Brannon, 1989; Miller, et al., 1980). These studies indicate that reliability varies across scales. However, in research relevant to the present study, Geyer et al. (1989) found alpha coefficients for GATB aptitude ratings greater than .80 when four raters were used (Table 10 summarizes their findings), suggesting that the aptitude ratings available from the *DOT* are also likely to be quite reliable.

The regression-estimated data are weighted linear composites of seven principal components comprising *DOT* ratings. Therefore, these composite scores should have higher reliability than the individual ratings (Li, Rosenthal, & Rubin, 1996), which themselves generally seem to have acceptable levels of reliability (with some exceptions; e.g., see Cain & Green, 1983).

Table 4:
Actual Test Score Data (DOT Level): Descriptive Statistics and Intercorrelations

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	.90	--						
3. Spatial Ability	102.42	10.93	.76	.84	--					
4. Form Perception	103.89	10.86	.77	.78	.78	--				
5. Clerical Ability	105.93	11.45	.81	.72	.56	.85	--			
6. Motor Coordination	102.39	9.16	.72	.64	.46	.79	.87	--		
7. Finger Dexterity	97.94	8.05	.43	.46	.46	.61	.42	.58	--	
8. Manual Dexterity	105.13	8.17	.32	.39	.45	.57	.43	.53	.60	--

Note. N = 518. All correlations are significant at p < .01.

Table 5:
Analyst Data (DOT Level): Descriptive Statistics and Intercorrelations

GATB Aptitude	*M*	*SD*	1	2	3	4	5	6	7	8
1. Verbal Ability	3.21	.80	--							
2. Numerical Ability	3.38	.81	.71	--						
3. Spatial Ability	3.30	.82	.35	.48	--					
4. Form Perception	3.23	.74	.26	.40	.70	--				
5. Clerical Ability	3.61	.83	.63	.59	.13	.16	--			
6. Motor Coordination	3.28	.60	-.08	.00	.35	.40	.01	--		
7. Finger Dexterity	3.30	.65	.00	.10	.37	.46	.05	.62	--	
8. Manual Dexterity	3.14	.58	-.21	-.10	.38	.34	-.24	.56	.50	--

Note. N = 518. |r| > .09 are significant at p < .05. |r| > .12 are significant at p < .01.

Table 6:
Regression Estimated Data (DOT Level): Descriptive Statistics and Intercorrelations

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	.97	--						
3. Spatial Ability	102.65	8.43	.84	.94	--					
4. Form Perception	104.08	7.10	.97	.98	.90	--				
5. Clerical Ability	106.06	8.02	.97	.92	.74	.95	--			
6. Motor Coordination	102.60	5.88	.94	.85	.64	.90	.97	--		
7. Finger Dexterity	98.15	3.10	.79	.77	.74	.86	.77	.81	--	
8. Manual Dexterity	105.21	2.47	.75	.81	.88	.86	.70	.67	.91	--

Note. N = 518. All correlations are significant at p < .01

Table 7:
Actual Test Score Data (SOC Level): Descriptive Statistics and Intercorrelations

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	.89	--						
3. Spatial Ability	105.59	10.05	.70	.81	--					
4. Form Perception	106.51	10.34	.78	.80	.76	--				
5. Clerical Ability	108.87	11.21	.84	.74	.53	.87	--			
6. Motor Coordination	104.23	8.53	.77	.66	.45	.79	.88	--		
7. Finger Dexterity	98.20	7.51	.51	.55	.57	.69	.51	.62	--	
8. Manual Dexterity	105.40	7.44	.34	.44	.52	.57	.43	.50	.55	--

Note. N = 264. All correlations are significant at p < .01.

Table 8:
Analyst Data (SOC Level): Descriptive Statistics and Intercorrelations

GATB Aptitude	*M*	*SD*	1	2	3	4	5	6	7	8
1. Verbal Ability	2.94	.81	--							
2. Numerical Ability	3.16	.76	.65	--						
3. Spatial Ability	3.15	.84	.19	.39	--					
4. Form Perception	3.14	.76	.15	.37	.72	--				
5. Clerical Ability	3.40	.76	.56	.47	-.10	-.03	--			
6. Motor Coordination	3.28	.60	-.21	-.08	.40	.43	-.10	--		
7. Finger Dexterity	3.31	.66	-.04	.08	.41	.51	.00	.68	--	
8. Manual Dexterity	3.17	.63	-.30	-.19	.42	.42	-.35	.67	.61	--

Note. N = 264. |r| > .14 are significant at p < .05. |r| > .18 are significant at p < .01.

Table 9:
Regression Estimated Data (SOC Level): Descriptive Statistics and Intercorrelations

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	.96	--						
3. Spatial Ability	105.35	7.36	.77	.89	--					
4. Form Perception	106.47	6.67	.97	.97	.85	--				
5. Clerical Ability	108.80	8.00	.97	.91	.64	.94	--			
6. Motor Coordination	104.50	6.16	.94	.83	.55	.90	.97	--		
7. Finger Dexterity	99.04	3.18	.73	.70	.65	.82	.71	.78	--	
8. Manual Dexterity	105.88	2.33	.65	.72	.83	.80	.59	.59	.89	--

Note. N = 264. All correlations are significant at p < .01.

Table 10:
GATB Reliability Estimates from Geyer et al. (1989)

GATB Aptitude	α: One Rater	α: Four Raters
General Ability	.88	.97
Verbal Ability	.93	.98
Numerical Ability	.75	.92
Spatial Ability	.78	.93
Form Perception	.72	.91
Clerical Ability	.70	.90
Motor Coordination	.65	.88
Finger Dexterity	.68	.89
Manual Dexterity	.51	.81

Clustering Results

Number of clusters. As discussed previously, the CCC, pseudo *F*, and pseudo t^2 indices were jointly examined to determine the number of clusters present in each dataset. Unfortunately, as often happens in practice, these statistical indices did not converge onto the same number of clusters, and so no clear solution appeared for any of the datasets. Therefore, rather than arbitrarily focusing on a single solution based on weak evidence, multiple cluster solutions were explored for each dataset and used in subsequent analyses. First, it was assumed that given approximately 500 jobs at the *DOT* level and 250 jobs at the SOC level, somewhere between approximately 2 and 50 clusters would end up being the most useful, manageable, and appropriate for most practical purposes. Second, for several of the datasets, peaks in graphs of the three types of clustering statistics appeared in both the 2-5 cluster range and 15-25 cluster range. Therefore, choosing cluster solutions above and below the 15 cluster point seemed appropriate. Based on this, three ranges were chosen to reflect a low range (below 15 clusters), a middle range (a range of 20 solutions starting at 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 within each of these three ranges was chosen based on examining the CCC, pseudo *F*, and pseudo t^2 plots (see Table 11). 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 is still helpful for understanding how the three types of data behave across different clustering solutions.

Similarity of cluster solutions across data types. Similarity of cluster solutions can be examined in several ways, such as by cross-classifying solutions or computing various indices of similarity. Previous research (Milligan & Cooper, 1986) suggested examining similarity of cluster solutions resulting from the three types of data using the Hubert and Arabie (1985) adjusted Rand (1971) statistic, which indicates the extent to which pairs of jobs placed in the same cluster for one solution are also placed in the same cluster for the other solution (and, conversely, jobs placed in different clusters for one solution are also placed in different clusters for the other solution). To some extent, this sort of agreement can occur by chance, and the adjusted Rand statistic takes this into account, similar in spirit to the kappa coefficient of agreement. The adjusted Rand statistic has an upper bound of 1.0, indicating perfect agreement, and should equal 0.0 when cluster solution agreement can be attributed to chance; however, it can also take on negative values when agreement is less than that expected from chance. This statistic can be interpreted as the proportion of agreement between cluster solutions over and above the agreement expected due to chance (Collins & Dent, 1988).

Table 12 presents Hubert and Arabie adjusted Rand statistics for each cluster range at the *DOT* and SOC level. This table reveals a couple of patterns. First, the three types of data, though related, clearly produce substantially different cluster solutions. None of the adjusted Rand values are above .45 (i.e., 45% agreement over what is expected due to chance) and only two are above .20, indicating little cluster solution agreement above chance levels in most cases. Second, it appears that (a) job analyst and regression-estimated data tend to produce more similar cluster solutions than do (b) actual test scores and job analyst data or

Table 11:
Number of Clusters Indicated by the CCC, Pseudo F, and Pseudo t^2

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

Table 12:
Adjusted Rand Statistic

Comparison	2-14 Cluster Range	15-34 Cluster Range	35-54 Cluster Range
DOT Level			
Actual Test Score and Analyst	.19	.07	.04
Analyst and Regression-Estimated	.42	.16	.12
Actual Test Score and Regression-Estimated	.20	.06	.03
SOC Level			
Actual Test Score and Analyst	.11	.08	.07
Analyst and Regression-Estimated	.15	.18	.16
Actual Test Score and Regression-Estimated	.24	.07	.06

Note. Adjusted Rand statistic from Hubert and Arabie (1985).

(c) actual test scores 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 statistics than did the actual test score and job analyst comparison and the actual test score and regression-estimated comparison. However, although the job analyst and regression-estimated comparison tends to produce relatively higher values, these values still appear to be relatively small in an absolute sense: the highest of the six is .42, and the other five are within .10 to .20, indicating that, altogether, substantial disagreement exists across all cluster solutions.

Note that a comparison between these clustering results and an appropriate "gold standard" would provide further useful information. For instance, one could determine whether clusters produced by one of the data types tend to be more similar to an established classification system than clusters produced by other data types. For example, we might compare our clustering results with the widely used SOC classifications. However, classifications in the SOC are made based on "work performed, skills, education, training, and credentials," (Bureau of Labor Statistics, 2001a) rather than required abilities; therefore, observed differences between the ability-based classifications in this study and the classifications of this system may not be easily interpretable, as they may reflect legitimate differences rather than unwanted deviations from a gold standard. One could compare our results to an ability-based classification system; for instance, Gottfredson's Occupational Aptitude Patterns Map is based on jobs' overall aptitude level (5 levels) and aptitude patterns (4 general patterns; see Gottfredson, 1986). This possibility could be explored with our data, though our decision was to be informed by the relative comparisons of cluster solutions with our existing data sets and not to stray from the O*NET classification system.

Criterion-Related Validity Results

DOT level analyses. Criterion-related validity coefficients are available for each of the 518 *DOT* jobs included in this study, correlating the nine GATB aptitudes with ratings of either job or training performance, or course/exam grades. Table 13 presents descriptive statistics for these criterion-related validity coefficients.

For test-validation purposes, analyzing criterion-related validity coefficients highlights the advantages and disadvantages of using each of the three types of data in job clustering. Useful cluster solutions will contain similar profiles of criterion-related validity coefficients within clusters yet differ between clusters. Profile analysis provides one way to examine variability in criterion-related validity coefficients across clusters. Profile analysis is an application of multivariate analysis of variance (MANOVA) for investigating two or more groups that are measured on multiple dependent variables (DVs). Profiles of scores can consist of either one DV measured multiple times or of different DVs measured on the same scale. In this case, the groups are job clusters, and the DVs are criterion-related validity coefficients for each of the nine GATB scales. Across job clusters, this method independently tests for differences in the *levels* of the aptitude profiles, differences in the *shapes* of the aptitude profiles, and a general test for the *flatness* of the pooled aptitude profile across all profiles (cf. Harris, 1975).

Table 13:
Criterion Related Validity Coefficient Descriptive Statistics at the DOT Level

GATB Aptitude	M	SD	Sample Size *M*	Sample Size *SD*
General Ability	.23	.16	82.0	68.4
Verbal Ability	.17	.16	88.6	66.2
Numerical Ability	.22	.16	88.6	66.2
Spatial Ability	.16	.16	88.6	66.2
Form Perception	.18	.16	88.6	66.2
Clerical Ability	.19	.16	88.6	66.2
Motor Coordination	.15	.16	86.1	64.2
Finger Dexterity	.15	.17	86.1	64.2
Manual Dexterity	.16	.18	86.1	64.2

Note. N = 518 for validity coefficients. Due to missing data, N = 388-518 for sample sizes.

The "levels" test assesses whether different clusters of jobs have different mean levels of criterion-related validity. The null hypothesis for this test is that the overall profile mean (i.e., the mean across the means of the separate DVs in the profile) is identical across groups (Harris, 1975). Rejecting this hypothesis indicates that job clusters differ in terms of mean DV levels, suggesting a main effect for cluster membership. Results from the present analysis would indicate whether clusters are significantly different in terms of their average level of criterion-related validity.

The "flatness" test assesses the null hypothesis that the pooled profile of DVs across all job clusters is flat (Harris, 1975). Results from the present analysis would indicate whether certain GATB scales tend to have more or less criterion-related validity than other scales. Thus, whereas the "levels" analysis tests for between-cluster differences based on mean validity coefficients across aptitudes, the "flatness" analysis tests for between-aptitude differences based on mean validity coefficients across clusters.

Finally, the "parallelism" test assesses the null hypothesis 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 groups differ significantly in terms of the shape of their DV profiles. Here, a significant interaction indicates that the job clusters' criterion-related validity profiles for the GATB scales are not parallel (i.e., their shape depends on the cluster). These three tests are therefore analogous to two-way analysis of variance tests. 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 (Harris, 1975).

Note that the following results are based on analyses that include criterion-related validity coefficients for G, which as mentioned previously is redundant with V, N, and S. Analyses excluding G were also conducted, yielding very similar results. Some minor differences were observed, particularly for the flatness analyses (e.g., in the magnitude of partial eta-squared values), but the pattern of results remained the same for all analyses. Furthermore, results from the flatness analyses are rendered somewhat irrelevant by the presence of significant parallelism results (discussed below); therefore the flatness results (and any differences in these results when G is excluded) are not of primary concern. Thus, results based on analyses including validity coefficients for G are reported for the sake of completeness.

Table 14 presents results of the "levels" tests. 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 ($p > .40$ 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 ($p > .20$). 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 middle 15-34 cluster range where there is some differentiation. Partial eta-squared values, indicating the proportion of variance in averaged validity coefficients accounted for by cluster membership, also demonstrate this pattern. 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. By contrast, in some cases, such as with actual test score data at the 35-54 cluster range, we obtained values as high as approximately 20%. In these cases it appears cluster membership is important, yet nontrivial within-cluster variability in average validity coefficients remains.

Table 14:
Profile Analysis "Levels" Test at the DOT Level

	Source	*df*	*F*	*p*	Partial Eta-Squared
2-14 Cluster Range					
Actual Test Score	3 Clusters	2	4.48	.01	.02
	Error	515			
Analyst	3 Clusters	2	0.67	.51	.00
	Error	515			
Regression-Estimated	3 Clusters	2	1.18	.31	.01
	Error	515			
15-34 Cluster Range					
Actual Test Score	18 Clusters	17	3.32	.00	.10
	Error	500			
Analyst	23 Clusters	22	0.97	.50	.04
	Error	495			
Regression-Estimated	23 Clusters	22	1.79	.02	.07
	Error	495			
35-54 Cluster Range					
Actual Test Score	42 Clusters	41	2.83	.00	.20
	Error	476			
Analyst	48 Clusters	47	0.83	.78	.08
	Error	470			
Regression-Estimated	50 Clusters	49	1.14	.25	.11
	Error	468			

Table 15 presents results of the "flatness" tests. These results clearly show that the mean validity profile across all clusters is significantly different from flat ($p < .01$). This indicates that one or more of the GATB scales tend to predict criteria differently than the other scales. Partial eta-squared values indicate that 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. However, these results are not explored further both because they are not the primary focus of this paper, and they are essentially rendered irrelevant by results discussed below.

Table 15:
Profile Analysis "Flatness" Test at the DOT Level

	Source	*df*	Wilks' Lambda	*F*	*p*	Partial Eta-Squared
2-14 Cluster Range						
Actual Test Score	GATB	8	.454	76.42	.00	.55
	Error	508				
Analyst	GATB	8	.425	85.85	.00	.58
	Error	508				
Regression-Estimated	GATB	8	.426	85.46	.00	.57
	Error	508				
15-34 Cluster Range						
Actual Test Score	GATB	8	.508	59.68	.00	.49
	Error	493				
Analyst	GATB	8	.486	64.44	.00	.51
	Error	488				
Regression-Estimated	GATB	8	.438	78.17	.00	.56
	Error	488				
35-54 Cluster Range						
Actual Test Score	GATB	8	.502	58.22	.00	.50
	Error	469				
Analyst	GATB	8	.502	57.51	.00	.50
	Error	463				
Regression-Estimated	GATB	8	.477	63.10	.00	.52
	Error	461				

Results of the "parallelism" tests indicate the level and flatness results must be qualified, because there is a significant GATB scale × 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 (see Table 16). In other words, each type of data produces at least some clusters that differ in terms of validity profile shape. Although partial eta-squared values are fairly similar for results for each of the three types of data, regression-estimated data produce the largest values across cluster solutions, followed by job analyst data, and then actual test score data. Overall, partial eta-squared values are relatively low, ranging from .05 (for actual test score data at the 2-14 and 15-34 cluster range) to .13 (for regression-estimated data at the 35-54 cluster range), indicating a considerable amount of within-cluster variance in profile shape.

SOC level analyses. Results obtained using SOC-level data are very similar to those obtained using *DOT*-level data and therefore are not presented in detail. The pattern of results is essentially the same, particularly for the levels and flatness tests. Results of the parallelism test are also similar in that a significant GATB scale × cluster membership interaction exists across all data types and ranges of cluster solutions ($p < .01$). However, 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 eta-squared values, followed by job

Table 16:
Profile Analysis "Parallelism" Test at the DOT Level

	Source	*df*	Wilks' Lambda	*F*	*p*	Partial Eta-Squared
2-14 Cluster Range						
Actual Test Score	GATB * 3 Clusters	16	.896	3.57	.00	.05
	Error	1016				
Analyst	GATB * 3 Clusters	16	.846	5.53	.00	.08
	Error	1016				
Regression-Estimated	GATB * 3 Clusters	16	.829	6.26	.00	.09
	Error	1016				
15-34 Cluster Range						
Actual Test Score	GATB * 18 Clusters	136	.643	1.66	.00	.05
	Error	3607				
Analyst	GATB * 23 Clusters	176	.604	1.45	.00	.06
	Error	3701				
Regression-Estimated	GATB * 23 Clusters	176	.569	1.63	.00	.07
	Error	3701				
35-54 Cluster Range						
Actual Test Score	GATB * 42 Clusters	328	.433	1.27	.00	.10
	Error	3706				
Analyst	GATB * 48 Clusters	376	.392	1.23	.00	.11
	Error	3674				
Regression-Estimated	GATB * 50 Clusters	392	.336	1.38	.00	.13
	Error	3662				

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.

Pay Data Results

DOT level analyses. Pay data were not available for *DOT* level jobs.

SOC level analyses. Pay data were obtained from the U.S. Bureau of Labor Statistics and represent each SOC job's median annual income for the year 2000. The Bureau of Labor Statistics collects these data through the Occupational Employment Statistics (OES) program, which involves a yearly mail survey designed to estimate employment and wages for various jobs (Bureau of Labor Statistics, 2001b). The wage data used in this study are based on a survey of approximately 800,000 establishments. Table 17 presents descriptive statistics for the data available for 260 of the 264 SOC jobs included in this study.

Intraclass correlation coefficients (ICCs) are one way to examine statistically the similarity of pay within job clusters versus between clusters. There are numerous versions of the ICC, but essentially these coefficients give the ratio of the variance of interest (often between group variance) over the total variance, which is 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. Job clusters developed for use in job evaluation should include less within cluster variance in pay rates compared with the variance in pay between clusters. Therefore, relatively larger ICC values could be taken as an indication that a given cluster solution is relatively more useful for job evaluation.

Table 18 presents ICC values and associated 95% confidence intervals (cf. Donner & Wells, 1986) for actual test score data, job analyst data, and regression-estimated data at all three cluster solution ranges. Results indicate the three data types perform fairly similarly with respect to the ICCs: actual test score and regression-estimated data tend to perform slightly better than analyst data, but the point estimates obtained are fairly similar, and the confidence intervals show substantial overlap across data types. Therefore, for the pay data, there are not significant differences in performance across types of ability data.

However, some differences exist 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 ICC values below .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 tend to be more useful for job evaluation purposes. Note that using a relatively large number of clusters does not defeat the purpose of clustering; even in these situations, the goal of grouping jobs into a smaller number of meaningful groups can be achieved.

Table 17:
Pay Data: Descriptive Statistics

N	260
M	$32,770.50
Mdn	$28,500.00
SD	$14,900.95
Minimum	$13,330.00
Maximum	$114,170.00

Table 18:
Pay Data: Intraclass Correlations

Cluster Range	Data Type	ICC	95% Confidence Intervals
2-14	Actual Test Score	.36	-.10 - .83
	Analyst	.29	-.13 - .71
	Regression-Estimated	.49	.07 - .91
15-34	Actual Test Score	.57	.40 - .74
	Analyst	.45	.26 - .65
	Regression-Estimated	.52	.34 - .70
35-54	Actual Test Score	.56	.41 - .71
	Analyst	.49	.34 - .64
	Regression-Estimated	.57	.43 - .72

Finally, cluster membership generally accounts for moderate amounts of variance in pay rates, with ICCs ranging from .29 to .57. This suggests that although cluster membership does account for a reasonable amount of variance, there remain nontrivial amounts of within cluster variability in pay rates, which provides opportunity for further investigation into the nature of the job clusters if clusters such as these are used in job evaluation situations.

Discussion

This study examined similarities and differences in job clusters that result from using different types of job data in the clustering procedure, even when the data claim to measure the same constructs. Previous research has indicated that different types of job characteristic data from *different* psychological domains can produce substantially different job clusters (e.g., Cornelius et al., 1979; Ghiselli, 1966; Pearlman, 1980). This study took a different approach by examining the effects on job clustering of different types of data that came from within the *same* psychological domain of aptitudes yet use different measurement methods: employee test scores, analyst ratings, and regression estimates. We then systematically examined the implications of these effects for a few major purposes to which clusters are put in organizational research and practice: test validation, job evaluation, and career exploration.

Findings

Aptitude intercorrelations. Before going into the specifics of the clustering results and their implications, a general characteristic of the GATB data should be discussed, namely the level of aptitude intercorrelation. Aptitude measures are known to show positive manifold, or positive correlations (Carroll, 1993). The eight GATB dimensions used in this study are no exception, showing high levels of intercorrelation at both the *DOT* and SOC level, particularly for regression-estimated and actual test score data (see Tables 4-9). For example, actual test score data demonstrated an average aptitude intercorrelation of .63 at the *DOT* level and .65 at the SOC level; regression-estimated data demonstrated an average of .85 at the *DOT* level and .80 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 at the *DOT* level and .24 at the SOC level. The high level of correlation for actual test score and regression-estimated data likely reflects a general cognitive ability factor (*g*) measured by all tests requiring cognitive effort. It appears that the eight GATB dimensions, rather than measuring distinct attributes, are to some extent measuring the same attribute (general cognitive ability), particularly when measured by test scores and regression-estimated scores.

This high level of aptitude intercorrelation is important because it likely restricts the underlying cluster structure in terms of both the number of distinct clusters that can exist and the manner in which jobs can be grouped. If the different GATB aptitudes all tap general cognitive ability, then to some extent jobs can be differentiated or grouped primarily according to this single ability factor. At the same time, aggregating GATB aptitude scores within jobs may help make the unique aspects of each aptitude more reliable, which would make the profile shape information a useful supplement to the information provided by general cognitive ability or by the general level of the aptitude profile.

The positive intercorrelations between aptitudes appears to be less of an issue when analyst data are used to describe occupational ability requirements. It may be that although the GATB subtests actually measure *g* primarily, analysts may be making distinctions among aptitudes to a greater extent. In this case, it is possible to develop job clusters according to similarity in GATB profile patterns/shapes. Analyst data may in fact be more appropriate than mean test score data whenever the goal is to differentiate among or group jobs according to aptitude profile patterns or shapes (e.g., when matching individuals to jobs according to their strengths and weaknesses). The following sections describe the main findings of this study regarding the clustering results along with their implications.

Clustering results. The clustering results suggest two reasonable general conclusions. First, there is no clear number of job clusters underlying the actual test score, analyst, or regression-estimated data at the *DOT* or SOC level. The reason for this is unclear, but it may be that no "true" cluster structure exists. Aptitude requirements for the jobs 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 reflecting 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, we emphasize that in practice, job clusters are formed and used in a variety of job classification systems – whether or not such systems "carve nature at its joints." 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.

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, the indices used to detect the number of clusters may 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. However, other analyses of the data (which are not presented here) using different clustering methods and different indices for determining the appropriate number of job clusters also failed to converge on a clear cluster structure. Although this certainly does not rule out the possibility that a clear structure is present or that a particular method might have uncovered such a structure, it does suggest similar results might have been obtained had we focused on using other methods.

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 cluster ranges and at both the *DOT* and SOC level, there appeared to be relatively little cluster solution agreement across the data types above chance levels, with actual test score data tending to produce the most dissimilar solutions. This finding extends previous research indicating that different types of job data often result in substantially different job clusters. Previous findings suggested that choosing the psychological domain according to the purpose of clustering is essential when developing data for clustering. The current findings suggest that, even within a given psychological domain, the type of data (which result from different measurement processes) is also an important influence on job clusters. Because both the choice of psychological domain and type of data within a given domain can substantially influence job clustering results, it is prudent to consider both when developing data to be used in job clustering. The following sections discuss this further.

*Implications for career exploration and O*NET's Ability Profiler.* As noted previously, O*NET (the U.S. Department of Labor's computerized occupational information tool developed to replace the *DOT*) includes a career exploration tool called the Ability Profiler. For each *DOT* job, developers of this tool generated regression-estimated ability scores, representing predicted mean employee GATB test scores. Clustering results from this study suggested that profiles of mean employee GATB test scores are substantially different from regression-estimated ability profiles. As discussed above, many of the jobs considered similar (i.e., belonging to the same cluster) when described by actual test score data were considered dissimilar (i.e., belong to different clusters) when described by regression-estimated data. This suggests that 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. This does not necessarily mean that the Profiler is producing inappropriate occupational suggestions, but it does indicate that the recommendations currently produced by this tool may be data-dependent to some extent. In other words, the same jobs 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 these profiles. First, the level analyses seemed to indicate that actual test score data tend to perform better than analyst or regression-estimated data in terms of producing clusters that differ in overall validity profile level. However, the differences were relatively small and should not be overemphasized. In addition, the effect sizes obtained for this effect (ranging from approximately .00 to .20) were relatively small, indicating the presence of nontrivial amounts of within-cluster variance. Second, the flatness analyses clearly demonstrated statistically significant differences in mean criterion-related validities across GATB scales, indicating that one or more of the GATB scales tend to be more or less predictive of criteria than the other scales.

Parallelism tests, which qualify the previous results, indicated that at both the *DOT* and SOC level and for all three cluster ranges, each type of data produced clusters that differ significantly in terms of the shape of their criterion-related validity profiles. 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 job-analyst clusters; shape differences in job-analyst clusters, in turn, accounted for more variance than did shape differences in actual test score clusters. However, these differences were small, and a consistent pattern did not hold at the SOC level. 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.

In addition, although significant results were obtained for all parallelism analyses, shape differences accounted for relatively small amounts of variance. For example, partial eta-squared values varied from .05 to .13 for these effects. Although in some contexts these values may be acceptable, they could be viewed as relatively small for the present purposes, indicating 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 conversely 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 some inappropriate conclusions in test validation situations. Note that this outcome may at least partly reflect the difficulty encountered in finding clear cluster solutions; any cluster solutions obtained from the datasets used in this study may have been relatively "artificial." Again, however, situations where a cluster structure is imposed is not that uncommon in practice.

Implications for job evaluation. Overall, the 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 findings suggest that 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). Also, cluster membership generally accounted for moderate amounts of pay rate variability. ICCs ranged from .29 to .57, indicating the presence of nontrivial amounts of within-cluster variance. 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 inherently has superior qualities or is more useful overall. More importantly, the relative strengths and weaknesses of each type of data do not appear to impact job evaluation differentially, 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, suggesting that ability-based job clusters are useful in job evaluation situations, but should be used as just one part of the larger evaluation process. For example, ability clusters could be used to categorize and evaluate jobs initially. Then, other data and considerations could be used to further categorize and evaluate jobs to establish appropriate pay rates (e.g., specific tasks performed, labor market supply and demand).

Conclusions

The purpose of this study was to examine similarities and differences in job clusters produced by actual ability test scores, job analyst ratings, and regression-estimated ability requirements. Results indicated that these three types of data, though they measure the same aptitude constructs, produced substantially different job clusters. Although these differences did not appear to have clear implications for the practical purposes examined, results in the present study suggest that the type of ability data describing jobs influences resulting job clusters and, potentially, the organizational decisions based on them. Combined with past research findings, we can conclude that both the method used to measure job characteristics as well as the psychological domain to which the job characteristics belong have an important influence on job clustering.

References

1. Arvey, R. D., & Mossholder, K. M. (1977). A proposed methodology for determining similarities and differences among jobs. Personnel Psychology, 30, 363-374.
2. Baughman, W. A., Norris, D. G., Cooke, A. E., Peterson, N. G., & Mumford, M. D. (1999). Occupational classification: Using basic and cross-functional skills and generalized work activities to create job families. In N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jeanneret, & E. A. Fleishman (Eds.), An occupational information system for the 21st century: The development of O*NET (pp. 259-271). Washington, DC: American Psychological Association.
3. Bureau of Labor Statistics (2001a). Standard Occupational Classification (SOC) user guide. Retrieved May 1, 2003, from http://www.bls.gov/soc/socguide.htm
4. Bureau of Labor Statistics. (2001b). Technical notes for 2000 OES estimates. Retrieved October 11, 2002, from http://www.bls.gov/oes/2000/oestec2000.htm
5. 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.
6. Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3, 1-27.
7. Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. New York: Cambridge University Press.

8. Colihan, J., & Burger, G. K. (1995). Constructing job families: An analysis of quantitative techniques used for grouping jobs. Personnel Psychology, 48, 563-586.
9. Collins, L. M., & Dent, C. W. (1988). Omega: A general formulation of the Rand index of cluster recovery suitable for non-disjoint solutions. Multivariate Behavioral Research, 23, 231-242.
10. Cornelius, E. T., Carron, T. J., & Collins, M. N. (1979). Job analysis models and job classification. Personnel Psychology, 32, 693-708.
11. 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.
12. Donner A., & Wells G. (1986). A comparison of confidence interval methods for the intraclass correlation coefficient. Biometrics, 42, 401-412.
13. Duda, R. O., & Hart, P. E. (1973). Pattern classification and scene analysis. New York: Wiley.
14. Dye, D., & Silver, M. (1999). The origins of O*NET. In N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jeanneret, & E. A. Fleishman (Eds.), An occupational information system for the 21st century: The development of O*NET (pp. 9-19). Washington, DC: American Psychological Association.
15. Garwood, M. K., Anderson, L. E., & Greengart, B. J. (1991). Determining job groups: Application of hierarchical agglomerative cluster analysis in different job analysis situations. Personnel Psychology, 44, 743-762.
16. 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.
17. Ghiselli, E. E. (1966). The validity of occupational aptitude tests. New York: Wiley.
18. Gottfredson, L. S. (1986). Occupational Aptitudes Patterns Map: Development and implications for a theory of job aptitude requirements [Monograph]. Journal of Vocational Behavior, 29, 254-291.
19. Harris, R. J. (1975). A primer of multivariate statistics. New York: Academic Press.
20. Hartman, E. A., Mumford, M. D., & Mueller, S. (1992). Validity of job classifications: An examination of alternative indictors. Human Performance, 5, 191-211.
21. Harvey, R. J. (1986). Quantitative approaches to job classification: A review and critique. Personnel Psychology, 39, 267-289.
22. Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2, 193-218.
23. Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. Psychological Methods, 1, 98-107.
24. 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.
25. Milkovich, G. T., & Newman, J. M. (1990). Compensation. Homewood, IL: BPI Irwin.
26. 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.
27. Milligan, G. W. (1981). A review of Monte Carlo tests of cluster analysis. Multivariate Behavioral Research, 16, 379-407.
28. 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.
29. 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.
30. Milligan, G. W., & Cooper, M. C. (1987). Methodology review: Clustering methods. Applied Psychological Measurement, 11, 329-354.

31. Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. Journal of Applied Psychology, 82, 627-655.
32. Pearlman, K. (1980). Job families: A review and discussion of their implications for personnel selection. Psychological Bulletin, 87, 1-28.
33. 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.
34. Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66, 846-850.
35. Sarle, W. S. (1983). Cubic clustering criterion (SAS Tech. Rep. No. A-108). Cary, NC: SAS Institute Inc.
36. SAS Institute (1999). SAS/STAT user's guide, version 8. Cary, NC: Author.
37. Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.
38. U.S. Department of Labor. (1970). Manual for the USTES General Aptitude Test Battery. Washington, DC: U.S. Government Printing Office.
39. U.S. Department of Labor. (1979). Manual for the USES General Aptitude Test Battery. Washington, DC: U.S. Government Printing Office.
40. U.S. Department of Labor. (1991). The revised handbook for analyzing jobs. Washington, DC: U.S. Government Printing Office.
41. Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58, 236-244.
42. 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.
43. Wilk, S. L., & Sackett, P. R. (1996). Longitudinal analysis of ability-job complexity fit and job change. Personnel Psychology, 49, 937-967.

Endnote

[1] Two of the standard nine GATB aptitudes, General Intelligence and Numerical Ability, are excluded from Ability Profiler analyses because General Intelligence is redundant with Vocabulary, Numerical Ability, and Spatial Ability, and Numerical Ability is split into its two component tests, Arithmetic Reasoning and Computation (McCloy, Campbell, & Oswald, 1999).