



Extending Test Validation Results from One Plant Location to Another: Application of Transportability Evidence

Mark A. Johnson
James P. Jolly
Idaho State University

ABSTRACT

The EEOC Uniform Guidelines for Employee Selection Procedures require validation evidence for tests having adverse impact. The Guidelines identify criterion-related, content, and construct validity strategies as viable methods for establishing validity. The Guidelines also address the acceptability of other forms of validity, specifically transportability and validity generalization. Unfortunately, neither the guidelines nor the selection literature provide very much guidance concerning acceptable means for establishing transportability. This paper raises a number of questions facing researchers interested in establishing transportability evidence and describes how the transportability of validity evidence was implemented for a multi-plant food processing organization.

The purpose of this paper is to describe some of the processes used to establish the transportability of validity evidence from one plant to another. The EEOC Uniform Guidelines on Employee Selection Procedures identify methods for establishing test validation and for transporting (extrapolating) validation results from one facility to another. However, the guidelines provide little guidance on how transportability is to occur and the literature fails to provide clear guidance on how to conduct a transportability study. This paper presents some of the limited information the authors found on the transportability process and describes how we applied theory to establish transportability in a multi-plant organizational setting. We believe this information will prove useful to others who may wish to try implementing a similar validation approach.

Background

The authors conducted a test validation study for a battery of tests to be used at one of the plants (Plant A) of a multi-plant food-processing organization. Extensive job analyses were conducted. The researchers spent over a hundred hours observing workers perform their duties, and talked with workers, foremen, supervisors and managers. In addition, many workers completed tailor-made job analysis questionnaires (JAQs). Later, task inventories (TIs) were developed for two categories of jobs and were administered to samples of workers performing various production, packaging, and warehouse jobs. (Job titles included Operators, Operator Helpers, Packers, Laborers, Quality Inspectors, Material Handlers, and Forklift Truck Operators.)

After completion of a concurrent validation study at Plant A, the sponsoring organization requested that we validate the battery at three other plants. Rather than conduct separate validation studies at each of these three facilities or even perform a single multi-plant validation study including the three plants, we proposed the strategy of transporting validity evidence from



Plant A to each of the other three plants. Managers familiar with the multiple plants informed us that the work and jobs at the three plants were quite similar. We visited the other three plants and conducted job analyses by observing and interviewing workers. These analyses provided evidence of the similarities of jobs and work performed at the four plants. But was this job analysis evidence sufficient to establish the transportability of validity? Little information is available to answer this question. We took the position that since validation involves the “accumulation” of evidence, our client would best be served by our establishing a greater degree of such accumulated evidence rather than some minimal (or insufficient) amount. Consequently, we administered the task inventories and JAQs to workers at each of the other plants. These instruments, particularly the task inventories, provided us with quantitative measures to assess the similarities of work at the four plants. Our primary goal was to establish the similarity between plant A and each of the other three plants. It was not necessary to demonstrate the similarities across all plants—the validity study results were to be “transported” from plant A to each of the other plants. If the work at any one plant was similar to the work at Plant A, transportability was established and justified for that plant. This paper will focus on the data from Plant A and Plant B.

Job Analysis and Transportability

The foundation of any testing system development process is a thorough understanding of the nature of the jobs for which individuals are being selected. At Plant A we conducted very detailed job analyses which included employee questionnaires, interviews, task inventories, and over one hundred hours of observation of actual job duties.

Transportability is the process of taking criterion-related validation results from one organizational location and inferring that the validity evidence at the source location applies to the same or similar jobs at another organizational location (either of the same firm or another organization). To establish the similarity of jobs and associated knowledge, skill and ability levels (KSAs) between Plant A and the other facilities, we spent at least two person-days observing workers at each of the other facilities. We also interviewed managers who had worked in more than one of the facilities. We supplemented these data with the results of job analysis questionnaires and task inventories that were completed by cross-sectional samples of workers from each plant. All these results indicated substantial similarities in jobs and their associated, necessary KSAs across the plants.

At all plants, although we studied a variety of jobs, we found that to a large extent jobs have many commonalties. Even when two jobs are classified as separate pay grades, the reality is that the two jobs are more alike than different. Moreover, most workers serve as fill-ins for higher graded jobs. The differences, as few as they are, are more a function of job classification than actual task and duty differences. A key point for establishing transportability of the Plant A validation results to these other plants is that most of the jobs are more similar than different and, therefore, the required KSAs are identical to those for which we developed the testing battery. More importantly, the KSAs tested by the selection battery clearly are important at all four plants.



Relevant Literature

Section 1607.7 of the Uniform Guidelines for Employee Selection Procedures gives some general guidance for the use of validity studies not conducted by the user. *“Users may, under certain circumstances support the use of selection procedures by validity studies conducted by other users or conducted by test publishers or distributors and described in test manuals.”*

Section B of the Guidelines provides the conditions under which criterion-related validity evidence may be used from other sources. One of these conditions or requirements is *job similarity*. However, how can job similarity be operationalized?

Section 1607.7 C addresses multiunit studies: *“If validity evidence from a study covering more than one unit within an organization satisfies the requirements of section 14B [technical standards for validity studies] below, evidence of validity specific to each unit will not be required unless there are variables which are likely to affect validity significantly.”*

Similarly, Section 1607.15E(1) requires: *“A description of the important job behavior(s) of the user’s job and the basis on which the behaviors were determined to be important should be provided. A full description of the basis for determining that these important work behaviors are the same as those of the job in the original study (or studies) should be provided.”*

The American Psychological Association's Principles for the Validation and Use of Personnel Selection Procedures provide some guidance concerning the degree of job analysis required to establish job similarity for validity generalization (which involves a greater degree of generalization than transportability, but has not received sufficient support from the courts):

“A less detailed job analysis may be all that is required because there is so much previous job analytic work on the occupation of interest or because past research on the job allows the generation of sound hypotheses concerning predictors and criteria can be developed with little reference to a specific job analysis in a particular organization. When a systematic new job analysis is not completed, the researcher should compile reasonable evidence which establishes that the jobs in question are similar in terms of work behavior and/or required KSAs” (p.5).

The APA Principles (p.6) note that *“the amount of information required depends on the purposes to be served by the selection procedure.”* Accordingly, more detailed job analysis may be necessary if there is reason to *“question whether people with similar job titles are, in fact, doing similar work, or if there is a problem of grouping jobs with similar tasks or responsibilities than if the jobs can clearly be placed in homogeneous groups”*(p.6).

The APA Principles (p. 28) also qualify the use of validity generalization. *“The cumulative validity evidence may indicate generalizability of validity for a selection procedure only for particular kinds of jobs or job families. In such a case, reliance on validity generalization results for jobs in new settings or organizations should meet”* certain conditions including the following:



“The job in the new setting is similar to the job, or a member of the same job family, included in the validity generalization study. A wide variety of methods are available to examine inter-job similarity or family membership for validity generalization purposes. If sufficient information is available on the new job families to accurately assign it to a relevant occupational category, classification can be made without a formal job analysis.”

Hence, the scientific principles set forth by the APA may not even require job analysis. However, although the scientific and professional status of validity generalization has received wide support, the legal status is unclear.

In EEOC v. Atlas Paper Box Company, 868 F.2d 1487 (1989 US App) the Sixth Circuit Court of Appeals held that validity generalization ignores doctrines established by Albermarle vs. Moody, where the court held that “*a test may be used in jobs other than those for which it has been professionally validated only if there are no ‘significant differences’ between the studied and the unstudied jobs.*” (29CFR 1607.4(c)(2)) In Atlas, the employer simply assumed job similarity—an act which was insufficient in the eyes of the court; the absence of job analysis made the generalizability of tests inadequate.

Accordingly, successful applications of transportability evidence include an “adequate” job analysis. The job analysis must clearly document the similarities between or among jobs (Mahaffey, 1993).

In Bruckner v. Goodyear Tire and Rubber Co., 339 F.Supp. 1108 (N.D. Ala. 1972), support for the transportability of validity evidence was provided when the court “*found no significant difference between units and jobs at two plants.*” Similarly, Rivera v. City of Wichita Falls, 665 F.2d 531, 538 at n. 10 (5th Cir. 1982), provided support for transportability of validity involving police cadets and Friend v. Leidinger, 446 F.Supp. 361 (E.D. Va. 1977) and Brunet v. City of Columbus, 642 F.Supp. 1214 (S.D. OH), support the use of transportability evidence when testing firefighters.

However, in Dickerson v. U.S. Steel Corp., 472 F. Supp 1304 (E.E. 1978), the courts did not accept the adequacy of job analysis data to support transportability of evidence for craft apprentices from one firm’s plant to that of another company’s facility. The court held that there was “no basis” for concluding that there were no significant differences in the jobs at different locations (Mahaffey, 1993).

But what establishes that “no significant differences” exist? Is this criterion qualitative, quantitative, or both? Are statistical tests required? What statistical methods should be used? What levels of significance should be required? Can the failure to determine statistical significance be used to prove job similarity? These are just some of the specific questions with which we were faced as we attempted to establish transportability of our validity evidence from Plant A to the other plants. The literature gave us some limited guidance but for the most part we were on our own to determine how we would operationalize the requirement of establishing job similarity.



It appears that the key to applying validity generalization study results to another setting is the establishment of a “*linkage study, in a manner very similar to the transportability requirements of the Guidelines. Where a rational chain of evidence has been given, more often than not, positive results have been obtained*”(Mahaffey, 1993).

According to Arvey and Faley (1988), most courts rely heavily on the Uniform Guidelines in cases that attempt to use the results from other organizations’ validation studies to establish validity at their own location. The Guidelines state that users may rely on such studies only when there is evidence of the similarity of content and context across jobs. The Pegues v. Mississippi State Employment Service case (488 F.Supp. 239 (ND MS 1990)) is the only case where validity generalization research has been accepted without evidence establishing job similarity (Arvey & Faley, 1988). The Court in Atlas refused to acknowledge job similarity in the absence of job analyses. Arvey and Faley (1988) report that for the most part, the courts generally refer to, use, and support the Uniform Guidelines. Only time will tell us the extent (if any) that the courts will rely on Pegues for precedent.

Arvey and Faley identify several studies (Arvey and Mossholder, 1977; Arvey, Maxwell, & Mossholder, 1979; Stutzman, 1983, & Harvey, 1986) that identify and evaluate different procedures for establishing job similarity. Techniques such as ANOVA, MANOVA, and cluster analysis are recommended to establish if jobs differ and the degree of those differences, if any. These methods were used at a single location to establish the similarity of jobs within a single organization with the purpose of determining whether the jobs could be combined for a criterion-related validation study. The techniques could find applications for establishing the transportability of validity evidence but their data and statistical requirements make them impractical in most applied settings.

More questions than answers result from the existence of these methods concerning their possible application for establishing transportability: Are these methods required? Are these methods really useful? Do these tools of the scientific profession provide utility to the practitioner who is trying to balance the recommendations of the academic/profession and the legal requirements of the Guidelines and the Courts? Is there some middle ground?

In summary, the Guidelines state that job similarity must be established, and although the literature does present statistical methods that have been used to establish job similarity in a single location, such applications are impractical in most applied settings. Moreover, it is unclear whether these methods are acceptable from a legal perspective for establishing “no significant differences” among jobs, especially in multi-plant locations.

Data Analyses and Discussion

The purpose of this multi-plant study was to demonstrate the similarity of jobs, tasks and the corresponding required knowledge, skills and abilities (KSAs) at Plant A plant compared to Plant B. Statistical analyses were conducted to provide evidence of the appropriateness and success of these comparisons because validation evidence exists for Plant A. This demonstration of job similarity across plants provides evidence of the transportability of validity evidence to the other plants.



Task Inventory Data

Data were collected using a task inventory. Production and operating employees at Plant A (n=59) and Plant B (n=45) were asked to complete 123 task statements related to packaging and production jobs. The statements asked them to rate the frequency with which they performed a number of tasks and the importance of these tasks.

We performed a variety of analyses including descriptive statistics such as means and standard deviations for each plant. The average ratings on the frequency rating scale for the 123 task at Plant A and Plant B were 3.06 and 2.92 (Table 1), respectively, on a five-point scale. Results of a paired t- test of mean differences on the frequency scale between Plant A and Plant B were not statistically significant, indicating that the mean ratings on the frequency scale are the same at Plant A and Plant B.

TABLE 1

**Average Ratings on Frequency Rating Scale Across the 123
Task Inventory Statements and Paired t-tests of Means**

Average Ratings Across Task Statements FREQUENCY SCALE		p-value PAIRED T TEST OF
Plant A	Plant B	
3.06	2.92	0.1475

Inter-rater reliabilities for the frequency scale and the Spearman-Brown adjustments made for these correlations are .26 for both plants. These inter-rater reliabilities, although relatively low in terms of correlation values, are not low for our rating purposes. Interrater reliabilities above .50 are rare (Schmidt, Ones, & Hunter, 1992). In the present situation, given that the jobs included are not the same jobs, that the tasks performed by persons in the same job categories often vary, and that workers often rotate from job to job and from department to department, it is not surprising that the interrater reliabilities are .26. This notwithstanding, because we are dealing with mean ratings, the more appropriate correlational measure is the Spearman-Brown adjusted correlation. Application of the Spearman-Brown formula resulted in the following measures: Plant A, .95; Plant B, .94. These measures fall well above the levels typically cited as acceptable reliabilities (Gatewood & Feild, 1998).

The mean ratings on the five-point importance scale are 3.54 and 3.66 for the two plants (Table 2). The results of t-tests comparing the difference between means also indicate that the



mean ratings across the 123 matched task statements at the two plants are not statistically significantly different.

The inter-rater reliabilities for the ratings made on the importance scale are .22 and .19 and their corresponding Spearman-Brown Adjusted correlations are .94 and .91 for Plant A and B respectively.

The mean ratings made on the frequency scale and the importance scale were compared across the matched 123 task statements using a t- test (Table 3). The mean differences between the two scales (Frequency and Importance) were .48 at Plant A and .74 at Plant B, with both plants' means higher on the importance scale.

Correlational analyses comparing arrays of the mean ratings between Plant A and Plant B on the frequency rating scale produced an interplant correlation of .87. The interplant correlation of mean ratings of importance between Plant A and Plant B was .85.

TABLE 2

**Average Ratings on Importance Rating Scale Across the 123
Task Inventory Statements and Paired t-tests of Means**

Average Ratings Across Task Statements IMPORTANCE SCALE		p-value
Plant A	Plant B	PAIRED T TEST OF
3.54	3.66	0.160



TABLE 3

**Paired t-tests of the Differences Between Frequency and
Importance Scale Means at Each Plant on the
123 Task Inventory Statements**

	PLANT A AVERAGE	PLANT B AVERAGE	p-value PAIRED T TEST OF
FREQUENCY SCALE	3.06	2.92	0.000001
IMPORTANCE SCALE	3.54	3.66	0.000000

Correlations between the frequency and importance rating scales were obtained by correlating each individual's ratings on each of these two dimensions (intra-rater). The average intra-rater correlations were .74 and .71 for Plant A and Plant B, respectively. These results suggest that the raters tended to rate the frequency and the importance scales similarly. Raters appear to have differentiated between the frequency and importance of each task.

The means of frequency and importance task statements were computed across raters at each plant. The results indicate that at Plant A, 66 of the 123 task statements (54%) exceed the scale midpoint of 3.0 on the frequency scale and that 53 (43%) of these exceeded 3.0 for Plant B. Similarly, 110 (81%) of the task statement ratings on the importance scale at Plant A and 109 (89%) of those at Plant B exceeded the midpoint of 3.0 on the importance scale. This suggests that many of the task statements are applicable to both plants and that these tasks are performed frequently and are important to the jobs.

Statistical t-tests comparing each task statement's means from Plant A to Plant B on the frequency scale were performed. Eighty percent of the statistical t-tests comparing the frequency scale task statement ratings at both Plants A and B were not significant (98 of the 123). Similarly, statistical t-tests of the task statement ratings on the importance scale between Plants A and B were obtained. Here, 85 percent (104 of the 123 task statements) were rated the same (not statistically different). Thus, the great majority of task statements at Plant B were rated on average the same (no mean differences) as the corresponding statements rated at the Plant A facility on both the frequency and the importance rating scales. These findings provide further



support for the similarity of tasks performed (frequently and/or importantly), and not performed (frequently and/or importantly) at the Plant A and Plant B.

Non-statistical comparisons were also made. Those task statements that were rated on average at 3.0 or greater on the frequency scale were identified. A subset of 50 task statements (40.6 percent) was rated at or greater than 3.0 at both Plant A and Plant B. Thus, workers from both plants frequently perform this subset of tasks. These pairings demonstrate the commonalties between the two plants on the frequency ratings for dozens of task statements. This demonstrates the similarity of many specific tasks across Plant A and Plant B despite the large variety of jobs and specific tasks, duties, and responsibilities performed by workers.

Considering that 80 percent of the tasks were not statistically different, that 50 of the tasks were rated on average at or above 3.0 on the frequency scale, and that the average frequency ratings at Plant A and Plant B correlated .87, substantial evidence supporting the hypothesis that no differences exist between the two plants has been provided.

Multivariate tests comparing the array of task inventory statements at Plant A to those at Plant B failed to produce statistically significant differences (Table 4). The multivariate tests considered all task statements simultaneously rather than performing many individual t-tests for each task statement. Thus, this more comprehensive and systematic measure failed to demonstrate any differences between Plant A's set of ratings and those of Plant B. The multivariate conclusion: There are no differences between the ratings made on the task inventory at Plant A and Plant B. However, we recognize that the results of the multivariate tests may not be meaningful because of the low statistical power associated with the test when a large number of task inventory statements are used (as in the present case). This notwithstanding, the bottom line is that our collective analyses, particularly the more detailed descriptive and inferential statistics described earlier in this paper, provide a degree of evidence far surpassing that which we believe is necessary to establish the similarities of work for transportability purposes.

TABLE 4

**Results of Multivariate Tests Comparing Task Statement Frequency
Ratings of Plant A and Plant B Across Individuals and Across
the 123 Task Statements**

Test	Value	F	Hypothesis df	Error df	Signif.
Pillai's Trace	0.987	1.17	67	1	0.640
Wilks' Lambda	0.013	1.17	67	1	0.640



Summary

While the Uniform Guidelines accept the concept of transportability of validity evidence from one company location to another, neither they nor the literature provide very helpful advice on how much evidence is required to establish transportability. We surveyed relevant case law and the scientific literature and decided that a combination of job analysis observation and task inventory data provided a reasonable amount of evidence. We conducted a number of statistical tests of the task inventory data in order to establish similarity of jobs across locations. The task inventory and job analysis questionnaire data demonstrate the high degree of similarity between the work performed at Plant A and each of the other plants. The variety of analyses consistently and collectively point to the similarity of work. We hope this approach may be useful to others interested in establishing transportability of validity evidence across locations.



References

- American Psychological Association, Division of Industrial-Organizational Psychology. Principles for the Validation and Use of Personnel Selection Procedures. College Park: Author, 1987.
- Arvey, R. D., Faley, R. H. Fairness in Selecting Employees, 2nd ed. Reading, Mass: Addison-Wesley, (1988).
- Arvey, R. D., Maxwell, S. E., & Mossholder, K. M. Even more ideas about methodologies for determining job differences and similarities. Personnel Psychology, 32, 1979.
- Arvey, R. D., & Mossholder, K. M. A Proper Methodology for Determining Similarities and Differences Among Jobs. Personnel Psychology, 30, 1977.
- Brunet v. City of Columbus, 642 F.Supp. 1214 (S.D. OH).
- Dickerson v. U.S. Steel Corp., 439 F. Supp 55 (E.D. PA 1977).
- EEOC v. Atlas Paper Box Company, 680 F.Supp. 1184 (E.D. TN 1987).
- EEOC v. Atlas Paper Box Company, 868 F.2d 1487 (1989 U.S. App.).
- Equal Employment Opportunity Commission. (1978). Uniform Guidelines on Employee Selection Procedures, Federal Register 43, no. 166, Washington DC.
- Friend v. Leidinger, 446 F.Supp. 361 (E.D. VA 1977).
- Gatewood, R. D., & Feild, H. S. Human Resource Selection, Fourth Ed. Fort Worth: The Dryden Press, 1998.
- Mahaffey, C. (1993). Validity Generalization: Will it become the fourth major line of validity evidence or a pipe dream? Glendale: Psychological Services, Inc.
- Pegues v. Mississippi State Employment Service, 488 F.Supp. 239 (N.D. MS 1980).
- Rivera v. City of Wichita Falls, 665 F.2d 531, 538 at n. 10 (5th Cir. 1982)
- Schmidt, F.L., Hunter, J.E. & Pearlman, K. Task differences as moderators of aptitude differences in selection: A red herring. Journal of Applied Psychology, 64 (6), 609-626.
- Schmidt, F. L. & D. Ones. (1992). Personnel Selection, Annual Review of Psychology, 43



Stutzman, T. M. Within Classification Job Differences, Personnel Psychology, 36, 1983.