Large Samples Are Better, Aren't They?:

Pooling Subjects Across Organizations

Jay A. Gandy

Michael A. McDaniel

U.S. Office of Personnel Management

Washington, DC

Presented at 94th Annual Convention of the American Psychological Association, August 26, 1986, Washington, DC as part of the symposium: New Examinations in Meta-Analysis, Hannah Rothstein Hirsh, Chair.

The opinions expressed herein are those of the authors and do not necessarily reflect the official policy of the U. S. Office of Personnel Management. The assistance of David Synk and David Swarthout of the Northern Test Development Field Center, Michigan Employment Security Commission, and U. S. Employment Service in providing data for this research is gratefully acknowledged.

Large Samples Are Better, Aren't They?:

Pooling Subjects Across Organizations

My purpose today is to present findings bearing on the methodological question of how samples are chosen and validity coefficients are computed for the validation of employment tests. Specifically, should predictor data and supervisory ratings on subjects in a given job group be pooled across organizations or employers, or should validity coefficients be computed separately by organization and then averaged under a validity generalization paradigm?

It is well established and accepted that larger samples yield more stable estimates of employment validity relative to smaller samples. The warnings of Schmidt, Hunter, and Urry (1976) with regard to statistical power and the devastating effect of sampling error in small-sample validity studies has largely been taken to heart and acted upon in recent test validation work. The increased interest and activity in consortium projects is in large part a direct reflection of this concern. Thus, it is also broadly accepted--but not well established--that pooling data on subjects across organizations and employers is the most appropriate means of obtaining large samples when, typically, each organization or employer has only a small number of potential subjects in the job group of interest.

Given what is known about rater errors (e.g., Latham & Wexley, 1981), the question arises as to whether systematic differences in rater behavior across organizations and employers may operate to lower validity coefficients computed on pooled groups. (The same question holds of course for pooling data across individual raters, but individual raters were not identified in the present research.) It would appear that to the extent that ratings from different sources are affected differentially by leniency, central tendency, or interpretation of successful performance, the correlation between test scores and ratings would be lower in a pooled group relative to the average correlation across the separate subsample groups. Moreover, to the extent that criterion ratings reflect relative comparisons between subjects, as opposed to scores against objective standards, pooling across raters and employers would be expected to yield lower correlations.

The hypothesis that subsample aggregation tends to yield lower validity coefficients was investigated in two phases. The first project (Gandy, 1986) was carried out as part of a reanalysis and extension of Hunter's (U.S. DOL, 1983) meta-analysis of GATB (U.S. DOL, 1970) validation data. For this project, there were 590 validity studies using supervisory appraisal criteria, with a total *N* of 56,593. Validities from single-employer studies were compared with those from multiple-employer studies. Substantial differences were found for jobs of relatively high complexity. The evidence for attributing the differences to the pooling of data across employers, however, was indirect. That is, with limited controls on other factors which might affect validity, the differences between mean validities for single-employer and multiple-employer studies might be due to other factors. This question led to a second project which permitted direct measurement of the aggregation effect, if any, by comparing validities on data from multiple employers computed in two ways: (a) separately by employer, then averaged across employers and (b) a single validity coefficient computed on the pooled sample.

I will briefly describe the findings from the intial project where pre-existing validities from single-employer and multiple-employer studies were compared, and then describe our recent findings where the raw data on subjects was analyzed in single-employer and multiple-employer forms for the same jobs.

Phase I: Comparisons of Existing Studies

In the initial project, Hunter's earlier conclusion that validity of ability tests is moderated by job complexity was further supported. Although cognitive tests have validity for all jobs, the level of validity decreases as job complexity decreases. Validity of psychomotor tests, on the other hand, decreases as job complexity increases. Thus, investigation of possible moderator effects of other variables, such as subsample aggregation effects, must control for level of job complexity. Otherwise, conclusions regarding possible moderator effects could be due to differing levels of job complexity in the sample of studies analyzed.

Chart 1 shows the mean validities of the cognitive (GATB G) test composite for single-employer and multiple-employer studies plotted as a function of job complexity. For lower complexity jobs, subsample aggregation apparently makes little or no difference in validity. For higher complexity jobs, however, differences are seen in favor of single-employer studies; and the differences are greater as job complexity increases. We hypothesize that the reason subsample aggregation has little effect with low complexity jobs is that standards for effective performance are relatively simple and straightforward, and performance is more observable for simpler jobs.

Chart 2 shows similar data for the GATB perceptual composite (aptitudes S, P, and Q). We note that valdities of perceptual tests do not vary monitonically with job complexity, but again we see substantially lower valdities in multiple-employer studies for jobs of higher complexity.

Chart 3 shows results for the psychomotor composite (GATB aptitudes K, F, and M). Here the higher validities are found for the less complex jobs. Again, for the least complex jobs, validities are virtually the same for single-employer and multiple-employer studies. For jobs of moderate complexity, validities are substantially higher for single-employer samples. For jobs of greatest complexity, validities are approximately the same for single-emplyer and multiple-employer samples. We hypothesize that the absence of a difference for high complexity jobs is due to the fact that psychomotor tests have little or no independent validity for high complexity jobs, and that the observed level of validity is due primarily to the correlation between the psychomotor tests and cognitive tests.

Phase II. Disaggregation of Pooled Samples

As previously mentioned, in the second phase of the research we compared validity for each aggregated sample with the average valdity across the component single-employer groups. Thus, identical test scores and criterion scores entered into the comparisons. Data for this project consisted of the original data from all validity studies conducted by USES since 1972. (Predictor and criterion scores on individuals prior to 1972 were not available.) Cases were retained for analysis according to the following rules:

1. All subjects were rated on the same supervisory appraisal criterion instrument which was scored in the same manner for all subjects, that is, ratings on five aspects of job performance, with the instrument readministered after several weeks and with the ratings summed across both administrations.

2. Data was available from multiple employers for the same job.

3. Ten or more subjects were available from each employer.

Data were then partitioned into two groups: jobs of moderate complexity and low complexity. Job complexity levels were based on the DOT job analysis data scale, that is, a rating of the extent to which the job requires interacting with data and cognitive demands. Unfortunately, the available studies included no jobs at the highest level of complexity on the DOT data scale, that is, jobs characterized by synthesizing requirements; and only one job was included at the second level which is called coordinating. This single job was dropped from the analysis. Jobs of moderate complexity included those dealing with analyzing, compiling, or computing; that is data scale levels 2, 3, and 4. Jobs of low complexity included those dealing with copying or comparing; that is, data levels 5 and 6. As shown on chart 4, the moderate job complexity group included data from 64 jobs, 607 employers, and 9,116 subjects. The low complexity category included 39 jobs, 314 employers, and 6,359 subjects.

The employer-based groups had mean sample sizes of 15 for the moderate complexity jobs and 20 for the low complexity jobs. As n sizes become increasingly small, the observed r's become statistically biased estimates of rho (Olkin and Pratt, 1958). Adjustments were made in validities based on the Olkin and Pratt formulations. Validities were then corrected individually for direct curtailment on the predictors and for average criterion reliability using the Pearlman, Schmidt, and Hunter (1980) estimated mean reliability of supervisory rating criteria (.60).

Chart 5 shows the key comparisons of interest for cognitive, perceptual, and psychomotor test composites for jobs grouped on job complexity. The pattern of results is as predicted for both sampling method and job complexity. For jobs of moderate complexity, mean subsample validities are higher on the average than validities computed for pooled subsamples. This is true across the different types of ability measures, with the difference averaging .024 across the composites. For lower complexity jobs, subsample aggregation appears to make little difference.

This pattern of results is consistent with that found in the first phase research where previously computed validities were compared for single-employer and multiple-employer samples. The results here differ, however, in two noteworthy respects. Mean validities for moderate and low complexity jobs are generally lower in this sample of studies than for the GATB data base as a whole; and the difference between single-employer and multiple-employer studies is smaller. For jobs of moderate complexity, the difference in mean validity in favor of single-employer samples averaged .125 across test composites in the initial research, compared to .024 in the present analyses which used a single set of data for the two conditions. A reasonable conjecture appears to be that there were undefined but systematic differences in the way the present data were obtained from employers.

Use of validities computed separately by employer is further supported by examination of the validity distributions for single-employer and multiple-employer conditions. As seen in chart 6, validities tend to be slightly more stable when computed on subsamples and then averaged. The observed standard deviation of the frequency weighted average subsample validities is smaller than the observed standard deviations of validities from the aggregated samples for most test composites. Exceptions are the psychomotor composite for moderate complexity jobs and the perceptual composite for lower complexity jobs. Since the same subjects make up each group, the relative differences seen here would also be reflected in residual standard deviations after removal of sampling error. Although the differences in validity variance between methods are not large, these findings indicate that nothing is lost, and some improvement in stability of results may be gained, by using a meta-analytic approach as opposed to the pooled-sample method in validation research.

Finally, following the presumption described earlier that differences between single-employer and multiple-employer validities are due to the effects of rater error, comparisons of the criterion distributions were made. For moderate complexity jobs, the mean criterion standard deviation for aggregated samples was 7.7, compared to 7.2 for the mean standard deviation of the averaged subsample criterion ratings. Similarly, for the lower complexity jobs, the corresponding mean criterion standard deviations were 7.8 for the pooled samples and 7.2 for the averaged subsample standard deviations. The slightly smaller mean standard deviation for the subsample criterion distributions indicates a degree of restriction in range on the criterion, which, other things being equal, would lead to smaller validities for the separate subsamples. Other things are not quite equal, however, as evidenced by the fact that subsample valdities are higher on the average. Thus, we observe that pooling across employers tends to increase variance of the criterion measure without increasing validity.

In conclusion, we make two further observations. First, in the absence of systematic rater error, validities computed from pooled ratings would be expected to be virtually identical to those computed from subsamples and averaged (except for the negative statistical bias in very small samples). Future research should focus on the extent of the decrement in validity to be expected under varying conditions, particularly for higher complexity jobs, which were not included in the present research. Secondly, the substantially lower validities found generally in the data base for which individual observations were available suggest that methodology differences other than subsample aggregation may be more important in moderating validities for jobs of similar complexity. We note that the single-employer studies analyzed in the initial research project had sample sizes averaging several times larger than the single-employer samples in the present research. The question arises as to whether more careful attention may have been given to considerations such as rater orientation and to gaining full cooperation in those companies with large numbers of participating employees. In any case, further research is needed on the relevance and effects of various methodological controls in studies using supervisory rating criteria.

References

Gandy, J. A. *Job complexity, aggregated subsamples, and aptitude test validity: Meta-Analysis of the GATB data base.* (Unpublished manuscript) Washington, DC: U.S. Office of Personnel Management.

Latham, G. P., & Wexley, K. N. (1981). *Increasing productivity through performance appraisal*. Reading, MA: Addison-Wesley.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, *29*, 201-211.

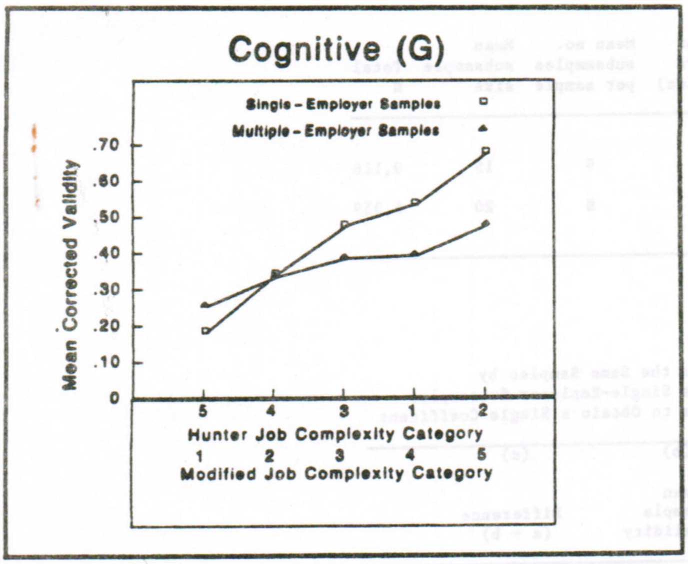
Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology*, *65*, 373-414.

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

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

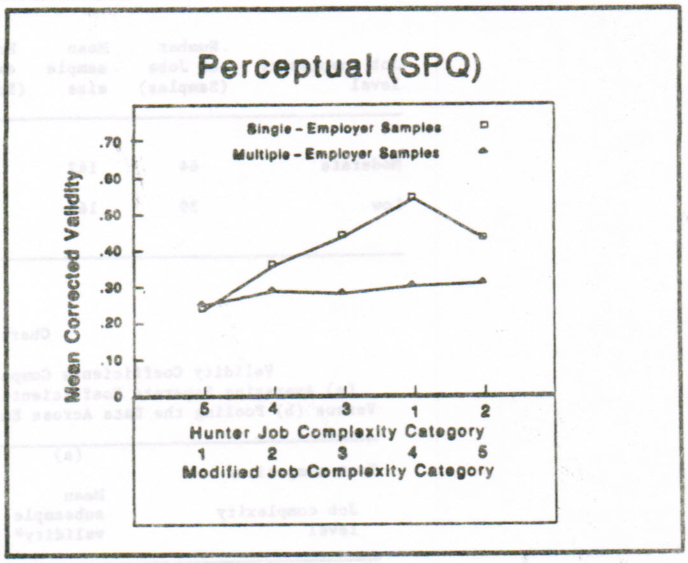
U. S. Department of Labor. (1983). *Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery*. USES Test Research Report No. 45; written under contract by J. E. Hunter. Washington, DC: Author.

Chart 1



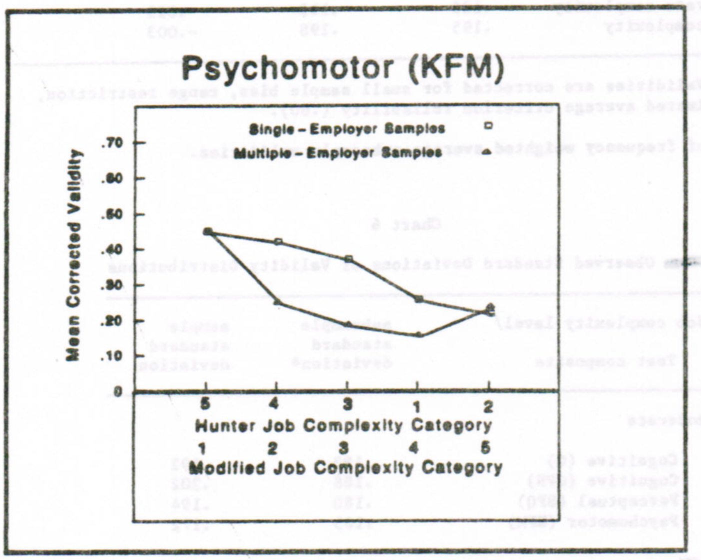
MEAN CORRECTED VALIDITY COEFFICIENTS FOR AGGREGATED (MULTIPLE-EMPLOYER) AND NONAGGREGATED (SINGLE-EMPLOYER) STUDIES AS A FUNCTION OF JOB COMPLEXITY.

Chart 2



MEAN CORRECTED VALIDITY COEFFICIENTS FOR AGGREGATED (MULTIPLE-EMPLOYER) AND NONAGGREGATED (SINGLE-EMPLOYER) STUDIES AS A FUNCTION OF JOB COMPLEXITY.

Chart 3



MEAN CORRECTED VALIDITY COEFFICIENTS FOR AGGREGATED (MULTIPLE-EMPLOYER) AND NONAGGREGATED (SINGLE-EMPLOYER) STUDIES AS A FUNCTION OF JOB COMPLEXITY.

Chart 4

Sample Characteristics

Job complexity level	Number of Jobs (Samples)	Mean sample size	Number of employers (Subsamples)	Mean no. subsamples per sample	Mean subsample size	Total N
Moderate	64	142	607	9	15	9,116
Low	39	163	314	8	20	6,359

Chart 5

Validity Coefficients Computed for the Same Samples by (a) Averaging Separate Coefficients Across Single-Employer Subsamples Versus (b) Pooling the Data Across Employers to Obtain a Single Coefficent

Test composite/ Job complexity level	(a) Mean subsample validity[a]	(b) Mean sample validity	(c) Difference (a - b)
Cognitive (G)			
Moderate complexity	.343	.321	.022
Low complexity	.310	.299	.011
Perceptual (SPQ)			
Moderate complexity	.264	.229	.034
Low complexity	.259	.252	.008
Psychomotor (KFM)			
Moderate complexity	.135	.112	.023
Low complexity	.195	.198	-.003

Note. Validities are corrected for small sample bias, range restriction, and estimated average criterion reliability (.60).

[a] Mean of frequency weighted average subsample validities.

Chart 6

[redacted] Observed Standard Deviations of Validity Distributions

Job complexity level/ Test composite	subsample standard deviation[a]	sample standard deviation
Moderate		
Cognitive (G)	.182	.192
Cognitive (GVN)	.188	.202
Perceptual (SPQ)	.180	.194
Psychomotor (KFM)	.183	.172
Low		
Cognitive (G)	.218	.222
Cognitive (GVN)	.202	.210
Perceptual (SPQ)	.189	.164
Psychomotor (KFM)	.142	.151

Note. Distributions are corrected validities.

[a] Standard deviation of frequency weighted subsample means.