ED 304 468 TM 012 843

AUTHOR MacLane, Charles N.; O'Leary, Brian S.

TITLE Job Specific Tests and an Overview of Research on Alternatives.

PUB DATE Aug 88
NOTE 15p.; Paper presented at the Annual Meeting of the American Psychological Association (96th, Atlanta, GA, August 12-16, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cognitive Tests; Employment Qualifications; Ethnic Groups; Factor Analysis; Government Employees; *Job Applicants; Job Skills; *Mathematics Tests; *Occupational Tests; Predictive Measurement; Reading Comprehension; *Reading Tests; *Test Construction; Test Items; Test Use; Vocational Evaluation; *Work Sample Tests
IDENTIFIERS Civil Service

ABSTRACT

The development of job-specific tests (JSTs) for two occupations is discussed. A reading comprehension test and a mathematical reasoning test were developed for Customs Inspectors, and a reading comprehension test was developed for Social Security Claims workers. JST items incorporated reading samples or math problems from those found on the job. Each job-specific reading test contained 40 items, and the Customs math test contained 30 items. Panels of subject matter experts rated tasks and test items. Correlational and factor analyses that related the two reading tests and the math test to cognitive or non-cognitive marker tests showed that the JSTs were cognitive tests that measured traditional verbal and mathematical abilities. Studies of the Customs tests with about 4,500 job applicants have confirmed the high reliabilities and generally good validities of the tests. The Claims worker test was not used operationally. Effect sizes for the Black (n=about 1,000) and Hispanic (n=about 1,000) Customs Inspector applicants were all close to one standard deviation with respect to the majority White group (n=about 2,500), which is typical of group differences associated with cognitive ability test scores. Research into alternative means of examining job applicants to reduce group differences indicated that combinations of interviews and tests, one of which should be a general cognitive ability test, can reduce group differences withou* losing test accuracy. Three tables present study data. (SLD)

Job Specific Tests and an Overview of Research on Alternatives

Charles N. MacLane
and
Brian S. O'Leary

U. S. Office of Personnel Management
Office of Personnel Research and Development
Washington, D. C. 20415

Paper presented as part of the symposium "The Federal Government's Search for Alternatives to Traditional Ability Tests" at the annual meeting of the American Psychological Association, Atlanta, Georgia, August 1988.

The opinions expressed in this paper are those of the authors and do not necessarily represent the opinions of the Office of Personnel Management.

Job Specific Tests and an Overview of Research on Alternatives

In this presentation, I will discuss something termed a "job specific" test and make some general remarks about the alternatives we studied. The device that we have called a "job specific" test is misnamed in that it is the minimum alterna:'ve: minimum in that, of the cognitively-oriented alternative tests, it's development involved the least replication of the job in the test. A reading comprehension test and a mathematical reasoning test were developed for Customs inspectors and a reading comprehension test was developed for Social Security claims workers. Customs inspectors do inspectional work in the enforcement of the Tariff Act and other laws governing the importation or exportation of merchandise. Claims workers adjudicate claims against the government by evaluating the legitimacy of an initial claim for retirement, disability, and/or health insurance benefits and by determining the amount of benefits to be paid initially and as the claim matures.

Job specific test items were written incorporating samples of reading materials or math problems selected representatively from those found in the job. A sample math item might ask Customs inspector applicants to pick, from multiple choices, the correct amount of duty to collect on 20 scarves worth $5.00 each when the specific duty rate is .16. To measure job-related reading skills, an applicant could be required to read a short paraphrased Customs or Social Security regulation and then pick the statement which is best supported by the paragraph. Table 1 in your handout shows examples of the kind of item which was developed for the Customs math and reading tests. The social security reading test was very similar in style to the Customs reading test.

In the development of the Customs tests, two panels of Customs subject matter experts (SME's) independently rated the learning and application of Customs laws and regulations and the collection of applicable duties and taxes as having "great importance" in Customs inspector work. To measure whether an applicant could perform these duties, a test of reading comprehension based on Customs-related laws and regulations and a test of mathematics reasoning based on the collection of duties and taxes were developed for the selection of Customs agents.

To begin development of the Social Security test, fifty claims SME's representing the various occupational series included in this type of social security work rated seven tasks relating to the learning and interpreting of social security rules and regulations as having high importance. The tasks were representative of the jobs found in the Claims area. A reading comprehension test based on randomly selected passages taken from social security rules and regulations manuals was developed. The process followed in the development of the Customs reading test included the following major steps: generating the essential reading list, determining the reading level of the job-related material, writing test items, and reviewing the test items. The source of the test items was a list of essential Customs inspector reading material that had been reviewed by a sample of entry-level inspectors and first line supervisors. The reading level for the job was calculated from the average scores for each book of reading materials (Payne, 1976). Then a panel of Customs inspectors was convened and given instructions on item writing by an OPM psychologist. The items were based on reading passages selected randomly from the essential reading materials.

The process followed in the development of the Customs math test paralleled that of the Customs reading test: initially, a group of job-related math-oriented materials was culled out by a panel of six Customs inspector SME's. The next step, the selection of math item types, did not have a reading test counterpart because math-related written material is reaaily converted to one particular reading test item type. The panel identified 16 tasks which were appropriate for testing. The panel also determined that two formats would be used for the items in the test: one type--the word problem--would present the required information in a narrative form, the second type--the table problem--would implant the data used to solve the problem among other data in a table or schedule.

The Claims reading test development began with the assembly of essential reading materials at job sites in three cities. A random sample of pages from these materials were selected for analysis of reading levels. Reading passages which fell within the average reading level for all the material were used as the basis for test items.

Each of the job specific reading tests contained 40 items. These tests were relatively easy. In the research samples, the mean of the Claims test was 30 (of 40 items) and the Customs reading test mean was 28. The Customs math test which had 30 items was more difficult with a mean of 17. The reliabilities were all in the .80's. Correlational and factor analyses which related the two reading tests and the math test to the cognitive and non-cognitive marker tests show that the job specific tests are cognitive ability tests which measure the traditional verbal and mathematical abilities which are the primary components of classic cognitive ability tests.

Concurrent criterion-related studies were carried out against training success and job performance. Training success was measured in Customs by classroom tests and in Social Security by ratings of training instructors. The performance rating measure duplicated in format the one used in studies of the other alternatives and it was used solely as a research instrument for which results were retained only in OPM files. Some of the dimensions which it measured varied with the occupations but many of the dimensions were identical to those measured in the studies of the other alternatives.

In general, the validity coefficients were typical of cognitive ability tests used for selection. The mean validity for all three tests against training criteria was .51 (corrected for unreliability), against job performance it was .37.

The best estimates of expected group differences on these measures are based on applicant data. Unfortunately, these are available only for the Customs tests because a decision was made on administrative grounds not to use the Claims test operationally. There have been about 1000 Hispanic and 1000 black applicants and about 2500 white applicants for Customs Inspector positions. The reliabilities of these tests are high and comparable and the sample sizes are relatively large so the estimates of groups differences should be fairly stable. The effect sizes for the black and Hispanic groups are all close to one standard deviation with respect to the majority white group. These estimates are close to those observed with the MT&E and the job knowledge test and are equivalent to the difference cited by researchers as being typical of group differences associated with cognitive ability test scores. Thus, the data on job specific tests do not support the hypothesis that building content validity into a cognitive test will reduce group differences. Validity, relative to cognitive ability tests in general, has been retained but so have the group differences. In sum, the job specific tests behaved as good cognitive tests should.

Initially I referred to the job specific test as the minimum alternative. In our studies, we wanted to see whether different forms of job specificity in test content and format could reduce group differences. The theory which led to this strategy is related to one of the five primary possible sources of test bias which Reynolds (1983) has outlined: although the points he made were couched in an educational context, it is useful to consider them because they reveal how thin our theorizing is in this area In paraphrase, they are (1) that the content of the tests is incompatible with the learning experiencies of minorities, (2) that the standardization samples of the tests don't include enough minorities, (3) that the language of the test is culturally alien, (4) that tests measure different attributes for different groups, and (5) that tests predict important criterion components differently or not all for minority members.

Of these arguments, the last is the only one which is completely compatible with the consistent finding that differential validity is a chance phenomenon (Bartlett, Bobko, Mosier, & Hannan, 1978; Hunter, Schmidt, and Hunter, 1979). That is, a test may be equally valid for the selection of members of all groups and still there may be the implication of unfairness in the selection process if one or more important criterion components are not predicted by the test and if these components may be predicted validly by another measure for which group differences are less. This reasoning leads, in its extreme form, to the Cosmic Search. The unreasonableness of the Cosmic Search comes about because it is difficult to find valia predictors of the job components which are not predicted by traditional cognitive tests and because we have no good theory of group differences in test scores so we don't know what to look for. (To say that group differences are due to differences in a general cognitive factor has not, by itself, led to many testable hypotheses for designing alternative tests).

We took the approach that if we developed measures which were more job specific than a traditional cognitive ability test (that is, more like the job in content or format), that we would be more likely to measure noncognitive components of the criterion or perhaps nontraditional cognitive components and that these measures might be valid and have smaller group differences.

Table 2 in the handout summarizes the results of the research studies we have been discussing. It shows the studies done for each procedure and summary and descriptive statistics for these studies. It is clear that the validities of these instruments are generally good, with the exception of the JCPS, and the E and E measures for which there was an inadequate data base. The validities for these measures are comparable to those reported for traditional cognitive ability tests. The descriptors (e.g., "good", "moderate") used to characterize the validities reflect both types of criteria and also reflect the level of corrections made to each statistic. This should be considered in making comparisons between procedures.

Secondly, factor analyses indicate that the MT&E, the job specific tests, and the job knowledge tests load heavily on a general cognitive factor and that these are the tests which show the largest effect sizes and the highest validities. (Only black-white differences are considered in these analyses.) The structured interview has a slightly lower overall validity, loads much less on the general cognitive factor, and has considerably lower effect size. The JCPS has little or no validity and very small effect sizes. The structured interview performed very well and seems to offer the best opportunity for reducing group differences. Before deciding that selections should be made on the basis of the interview alone, it should be remembered that the supervisory ratings used as criteria were collected for this research only. They would be freer from error than the typical ratings. More importantly, the structured interview was extensively and carefully developed with behavioral benchmarks to aid the raters' judgments. There were at least two raters, trained with videotapes produced for these studies, rating each candidate. Thus it is probable that the ceiling of the validity of the usual structured interview is lower than was observed in these studies.

If these conclusions concerning the structured interview are true, then the loss of validity by using it alone relative to a good cognitive ability test with a generalizable validity of over .50 would be considerable. An alternative is to use both a cognitive test and an interview. In order to estimate the validity and group differences when these instruments together for selection, an analysis was made of the MT&E and the structured interview as an equally weighted composite with a composite validity and and effect size. This analysis parallels one suggested by Schmidt (1988).

The basic data and results are shown in Table 3. The effect sizes of the two measures were estimated by cumulating across samples. Very small samples from some occupations were not included in the meta-analyses. The effect size of an equally weighted composite of the two instruments was estimated from the mean N-weighted cumulated effect size estimates. The validity of an equally weighted composite was estimated from the corrected estimates of the validities of the MT&E and the interview provided in the reports on these instruments. The results shown in Table 3 indicate that, even after correction for the composite unreliability, the effect size is .83. This is a reduction from the one standard deviation difference which has been our basis for comparison. The composite validity is .61. This validity could be even higher if regression weights were used. One caveat is that there was an unknown amount of indirect restriction in range on the interview scores. Comparison of the variances of the scores in the cumulated samples with other samples in which there should have been no restriction does not indicate that this should have been a problem.

These results support a strategy of test development which seeks to optimize combinations of tests, one of which should be a general cognitive ability test. There is obviously much work that can be done. It is very promising, however, that there appears to be a psychometric methodology which can reduce group differences in selection rates without lowering the accuracy of our tests.

This strategy does not relieve the test user of making utility decisions. The increased costs of administering alternative measures must be weighed against the probable decrease in adverse impact and increase in validity. The cost for the interview, for example, might be considerable.

REFERENCES

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. Personnel Psychology, 31, 233-241.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. Psychological Bulletin, 86, 721-735.

Payne, S. S. (1976). Reading ease level of D. C. Fire Department written materials required for entry-level job performance. Washington, D. C.: U.S. Office of Personnel Management.

Reynolds, C. R. (1983). Test bias: In God we trust; all others must have data. Journal of Special Education, 6(2), 311-313.

Schmidt, F. L. (1987). The problem of group differences in ability test scores in employment selection. Paper presented at the Conference, Fairness in Employment Testing. Sponsored by the Personnel Testing Council of Southern California.

U.S. Equal Employment Opportunity Commission, U.S. Civil Service Commission, U.S. Department of Labor, & U.S. Department of Justice. (1978). Uniform guidelines on employee selection procedures. Federal Register, 43(166), 38290-38309.

Table 1

Examples of Job Specific Test Items

Customs Math Item Example

Sample Question 2: **An importer has a shipment of 2,000 pens of equal value with a total value of $800.00. The duty rate on pens valued at 10¢ or more but not over 50¢ per pen is 8% of their value; the duty on pens valued over 50¢ but not more than $1.00 per pen is 6% of their value. How much duty is paid on the shipment of pens?**

A) $ 48.00
B) $ 64.00
C) $160.00
D) $180.00
E) None of these

Customs Reading Item Example

Sample Question 3.

When Congress passes a law, it does not include within the law details about how the law is to be administered. Therefore, for each law Congress authorizes, the department or agency that administers the law issues such rules and regulations as are necessary for its enforcement. The rules and regulations are usually published in proposed form in the *Federal Register* for public comment.

Select the statement that is best supported by the paragraph.

A) Public comment on laws proposed by Congress are published in the *Federal Register*.
B) The *Federal Register* must accept the rules and regulations that are published.
C) Congress empowers the agency that administers a law to set forth rules and regulations.
D) The legislative process may differ with different laws.
E) Congress establishes guidelines for enforcing the laws it passes.

Table 2

Summary of Alternatives Research

Alternative Procedure	Occupation(s) and Implementation	What It Measures	Validity	Total Sample Size	Impact	Comments
MT&E	-Tax Technician, Fall 1984 -Internal Revenue Officer Spring, 1986 -Social Security Claims Authorizer and Claims Representative, Winter, 1987 -Computer Specialist, Fall, 1982	Ability to learn the job related material required to perform in an entry-level position and progress to the journey level	Good Job Perf=.46[2] Training=.80[2]	 826 (perf.) 847 (trng.)	Large*	For entry-level positions; better for more structured jobs; past use for trades occupations
JCPS	-Computer Specialist, Fall, 1982 -Tax Technician, Fall, 1984	Compatibility between an applicant's preferences and special characteristics of job	Not useful Job Perf=.03 Training=-.04	 344 (perf.) 594 (trng.)	None*	
Job Knowledge Test	-Contract Specialist, Spring, 1986	Knowledge of Contract Specialist work	Good Job Perf=.38[1] Training=.59[1]	 393 (perf.) 410 (trng.)	Large*	
Structured Interview	-Tax Technician, Fall, 1984 -Internal Revenue Officer, Spring, 1986 -SSA C'aims Representative, Winter, 1987 -Customs Inspector, Fall, 1986 -Contract Specialist, Spring, 1986	Interpersonal "meet and deal" abilities	Moderate** Job Perf=.49[2] Training=.38[2]	 733 (perf.) 704 (trng.)	Small*	Mass screening of applicants difficult because of time and personnel required to administer the intervie
T & E Ratings	-Computer Specialist, Fall, 1982	Applicant's ability and motivation to perform job predicted from achievements and experiences	Undetermined[3] Job Perf=.04 Training=.38[2]	 162 (perf.) 218 (trng.)	Small*	Small sample make unstable esti mates of validity coefficients; this is our weakest database
Job Specific Test	-Customs Inspector, Fall, 1986 -SSA Claims Representative, not used operationally by agency request	Ability to understand job related math and reading materials	Good Job Perf=.37[1] Training=.51[1]	 600 (perf.) 498 (trng.)	Large*	Easier to develop than traditional ability test but has equivalent validity

Note: No overall unfairness (under the Cleary model) against minorities noted for any of the selection procedures.

*Adverse impact statistics based on the Uniform Guidelines (1978) 80% rule are unavailable because the small numbers of hires relative to the numbers of applicants makes these analyses unreliable. The statistic which is presented in this chart is effect size which is the difference between the mean scores for the majority group (white) and a minority group (here black only) divided by a measure of the variation of the scores. Cohen (1970) indicates that effect sizes of less than .20 show no impact, effect sizes of .20 to .50 are small, .50 to .80 are medium, and over .80 are large.

**Validity is expressed for the interview as a ranking procedure. Operationally, it was used as screen-out mechanism. Screen-out procedures, by definition, cannot be validated because there is no criterion data for those screened-out.

[1] Correlation derived from meta-analysis across samples and corrected for criterion unreliability.

[2] Correlation derived from meta-analysis across samples and corrected for criterion unreliability and range restriction.

[3] Validity coefficients for performance and training criteria are inconsistent, possibly because of small sample sizes (6 samples of 20-30 incumbents each). Overall validity not determinable from this database.

Table 3

Correlational and Effect Size Statistics for Estimating Composite Validity and Group Difference

Effect Size Statistics

Structured Interview

Occupation	Effect Size (d)	Black Group			White Group		
		Mean	SD	N	Mean	SD	N
Tax Technician	-.04	2.78	1.26	112	2.72	1.36	306
Internal Revenue Officer	.39	2.87	1.33	244	3.30	1.23	422
Claims Representative	.15	3.30	.88	40	3.44	.95	63
Contract Specialist	.20	3.33	1.06	83	3.55	.99	267
N-weighted d	.24						

MT&E

Occupation	Effect Size (d)	Black Group			White Group		
		Mean	SD	N	Mean	SD	N
Computer Specialist	1.10	45.33	15.29	2041	61.53	12.22	7672
Tax Technician	1.03	29.58	9.83	1210	40.92	9.06	1983
Internal Revenue Officer	.90	41.23	6.85	1291	47.72	5.30	2784
N-weighted d	1.01						

Mean weighted correlation of MT&E and Interview (corrected for unreliability) = .21

Mean weighted reliability of Interview = .91

Mean weighted reliability of MT&E = .92

validity of equally weighted composite of MT&E and Interview corrected for range restriction and unreliability = .61

Effect size of equally weighted composite of MT&E and Interview corrected for predictor unreliability = .83