DOCUMENT RESUME

ED 308 217 TM 013 560

AUTHOR Kuehn, Phyllis A.; And Others
TITLE Teacher Licensure Test Job Analysis Response by Gender, Race, and Age: Secondary Science and Mathematics.
PUB DATE Apr 89
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Age Differences; Cluster Analysis; Higher Education; *Job Analysis; Licensing Examinations (Professions); *Mathematics Teachers; *Racial Differences; *Science Teachers; Secondary Education; Secondary School Mathematics; Secondary School Science; Secondary School Teachers; *Sex Differences; Teacher Certification; Test Bias
IDENTIFIERS *Georgia Teacher Certification Testing Program

ABSTRACT

This study addressed issues raised in the literature on science and mathematics teacher certification testing concerning the validity of job analysis data and the test domain defined by the job analysis. More specifically, the issues addressed are those associated with race, gender, and age. Questionnaires were sent to 2,801 mathematics and 2,468 science teachers or teacher supervisors identified by the Georgia Department of Education as certified in these fields. A total of 25 different forms of the Georgia Teacher Certification Test had been developed to represent the fields of secondary certification (grades 7 through 12) in the state. The forms were distinguished by task statements pertinent to particular subjects taught; the science form had 148 unique task statements, while the mathematics form had 160 such statements. For science and mathematics teachers, respectively, 1,384 and 1,600 usable responses were available. Teachers rated the task statements by indicating for each one whether they actually performed the task, its importance to the learning process, and the possibility of successful performance by minimally competent teachers. Task statement content clusters were identified as well as simple effects and group mean differences. Results indicate significant effects based on race for science teachers, but no other significant effects based on race, sex, or age. Fourteen data tables and six graphs present study data. (TJH)

Teacher Licensure Test Job Analysis Response by Gender, Race, and Age: Secondary Science and Mathematics

Phyllis A. Kuehn
William M. Stallings
C.L. Holland
Georgia State University

Paper presented at March, 1989 AERA Annual Meeting
Revised April, 1989

Send Correspondence To:

Phyllis A. Kuehn, Ph.D.
Applied Linguistics and ESL
Georgia State University
Atlanta, GA 30303

Issues have been raised in the literature concerning the validity of job analysis data and the test domain defined by the job analysis. For example, it has been suggested that job content may not be independent of personal characteristics of job incumbents. In this work, teacher importance ratings of secondary science and mathematics job analysis task statements were evaluated for possible response differences by gender, race and age. The results showed that while content subareas in both fields were rated of differential importance by teachers, in general there were not important group rating differences that would have lead to gender, race, or age bias in a definition of the test content domain based on the job analysis responses.

Teacher Licensure Test Job Analysis Response by Gender, Race, and Age: Secondary Science and Mathematics

A significant aspect of the current reform movement in education is the increasing use of assessment procedures to evaluate prospective education students, education graduates seeking initial certification, and veteran certified teachers (Shulman, 1986; Mehrens, 1986 & 1987; Lehmann and Phillips, 1987). In different states, this has resulted in a variety of assessments, including basic skills, professional or pedagogical knowledge, general knowledge, subject-matter knowledge, and performance ratings. In some states, examinations are produced locally; more commonly, states adopt for use tests produced by the Educational Testing Service (ETS) and other testing concerns such as National Evaluation Systems (NES) (Lehmann and Phillips, 1987).

Fundamental to all test construction and use is the demonstration of test validity, or the correctness of the interpretation of test scores (Cronbach, 1971). For licensure or certification tests, the appropriate validity strategy is the content validity approach (Kane, et al. 1989), which requires the establishment of job-relatedness of the content domain to be sampled by the test (APA, AERA, NCME, 1985; Mehrens, 1986 & 1987). Except for tests developed nationally, such as the National Teachers Examination (NTE), and adopted after state-level validation studies are conducted (Cross, 1985), job-relatedness is generally based on the results of a job analysis that is conducted during test construction. Mehrens (1986) states that "What appears to be the most common and feasible approach for doing the job analysis is through a survey of the people in the profession" (p. 28).

The Guidelines (Uniform Guidelines on Employee Selection Procedures, EEOC, 1978), although specifying the necessity for job analysis in validation, state:

> Any method of job analysis may be used if it provides the information required for the specific validation strategy used. (p. 38300)

The Standards (11.1) (Standards for Educational and Psychological Testing, APA, 1985) also state "job analyses provide the basis for defining the content domain [in licensure and certification tests]" (p. 64), but do not specify any job analysis methodology. The other document that guides test standards, the APA Principles for the Validation and Use of Personnel Selection Procedures (1987) also provide no guidance in job analysis methodology or adequacy. For high-stakes tests such as the Georgia teacher certification tests (TCT), the decision of validity or invalidity is often resolved in court. The adequacy and validity of the job analysis is critical to the content validity argument and the inferences to be drawn from the test results (Elliot, 1987).

However, there is a dearth of research on job analysis in licensure testing, and few guidelines are available from personnel psychology or other literature with which to frame an evaluation of job analysis adequacy and validity.

An aspect of job analysis validity examined here involves investigating possible job analysis response differences associated with job incumbent characteristics (sex, race, age). The teaching fields considered were secondary science and mathematics. Both are subject-matter content intensive teaching fields, and both have a relatively high proportion of male teachers.

The Need For Job Analysis Evaluation

The need for job analysis evaluation is expressed in the Standards:

> Probable sources of variance that would confound the construct or domain definitions underlying the test should be investigated by the test developer, and the implications of the results for test design, interpretation, and use should be presented in the technical manual or in supplementary reports. (p. 28, Standard 3.12)

Tenopyr (1986), commenting on the proliferation of job analysis techniques, notes:

> Although a universal job analysis system is not advocated, it appears that there is a need for developing some principles for analyzing jobs. Despite the large number of job analyses being done today, there does not appear to be available the research base from which the needed principles can be drawn....The major question of the validity of the masses of data which have been generated is of utmost importance....Various types of raters should be examined, e. g., supervisors, incumbents, psychologists. Different specificity levels of construct should be employed. Studies to determine the degree of response style associated with such ratings should be undertaken. (p. 283)

Barrett (1981) points out simply that "there is no agreement on what makes an adequate job analysis" (p. 586).

Prien notes in his 1977 review that although job analyses have become increasingly important in selection procedures, there is still an "absence of research which defines the necessary and sufficient job analysis method" (p. 167). He identified basic issues in the context of content validity that require attention: 1) job analysis reliability and validity, 2) job functions and individual differences, 3) the research designs needed to produce appropriate information, and 4) the sufficiency of job analysis information.

Guion (1978) raises other issues regarding fairness and bias in the content domain. He states that "the idea that content domain samples are inherently fair seems widespread" and that this is probably a correct idea because "carefully constructed content domain samples seem likely to be free from bias" (p. 502). However, he warns that this "assumption of fairness may be vulnerable at several points," one of which is the assumption that the "job content domain is independent of the characteristics of the people who hold the job" (p. 502). This may not be true if "actual job content differs in different subgroups of incumbents" (p. 502). If it is the case that "important and testable aspects of what are actually two jobs [are] treated as one, then a test sampling of one is unfair to applicants for the other" (p. 503). This might be the case, according to Guion, when an affirmative action hiring produces "qualitatively different jobs for men and for women or for minority and nonminority employees" (p. 503).

Guion (1978) also speculates that job content might not be independent of personal characteristics in positions that allow different styles of work. In this case:

> Over time, these differences may produce qualitatively different jobs. This gradual process of change can be described as "drift" in job definition. Groups of people who think and behave alike and are in continuous communication with each other may drift in common ways. If these differences are identified with racial differences, a cultural drift may occur. In what is supposed to be an increasingly integrated society, we paradoxically find more and more voluntary social isolation of minorities. This sets a stage for drift in different directions for different cultural groups. A parallel drift that has no connotation of race or sex may occur for people who use different competencies in achieving the same ends. (p. 503)

If such drift is trivial, Guion goes on to say, it becomes a trivial issue in content domain definition. "However, a substantial problem of fairness could arise when the test content domain is substantially defined by purely stylistic elements irrelevant to the actual quality of performance on the job" (p. 503). It is certainly the case that competent teaching allows for diversity of style. Madaus (1987) also raises many issues about the current validation methods employed in licensure test validation. He questions the validity of the expert judgments used to define the test domain and challenges researchers to generate and test disconfirming hypotheses about the validity of the test validation methods currently employed.

In addition to these general questions about the validity of job analysis data, there are results of research on job ratings and other rating research indicating that specific response

differences have been observed or that response differences might be expected. In job analyses conducted to establish the job-relatedness of tests designed to measure pedagogical knowledge (Elliot, 1987) and non-subject-matter related job content (Potter, 1980), teacher rating responses to task statements were different for educators who differed on job environment (Elliot, 1987) or job context variables (Potter, 1980) such as grade levels and teaching environment. Outside variables such as experience and training background did not produce significant response differences.

Bemis, Belenky, and Soder (1983) note:

> While no particular job analysis method is prescribed, government regulations and court decisions clearly indicate a belief that job analysis will lead to fair job-related personnel practices. Very little research has been done on this belief or on other general job analysis questions. A preliminary study on this question conducted by ...[Boyles, Palmer, and Veres, 1980] indicates that blacks and whites may respond differently to judgmental job analysis questions. (p. 144)

Veres, Boyles, and Champion (1983) report that Boyles et al. (1980):

> Found significant differences between black and white subject matter experts (SMEs) on job analysis ratings of clerical jobs. In this case, differences in job analysis ratings were not accompanied by similar differences in scores on the selection device. Black applicants in fact scored highest (vis a vis White applicants) on those areas of the selection test that black incumbents had rated lower than white incumbents on the job analysis. (p. 3)

Veres et al. (1983) suggested that "these findings appear to preclude blind faith in the fairness assumption without further study" (p. 3).

Veres et al. (1983) pointed out that Boyles et al. (1980) did not determine whether the differences in ratings were the result of different job content or different response ratings of the same job content. In a similar study, Veres found "no evidence of [significant] racial differenes in rating accuracy," (p. 6) but found "a number of inaccurate raters within each racial group" (p. 6) rating real and bogus tasks associated with their own jobs. These differences were greatest in ratings of criticality, job-entry preparedness, and the relationship of task performance and overall job performance. The differences in the ratings of these variables could lead to selection devices that are biased.

Rationale for This Study

The reasons for evaluating the job analysis responses by race, gender, and age are as follows. First, in the fields of science and math, there is evidence of performance differences for male and female students on standardized tests. It is possible that science and math TCT content knowledge exams could result in adverse impact for female job applicants or current teachers. Pallas and Alexander (1983) state that it is:

> Well-established that by the end of high school there are large sex differences [favoring boys] in mathematical aptitude and achievement...[and that this difference is] recurrent across countries and over time, and is obtained on various standardized tests. (p. 165)

The authors cite mean differences of 41 to 51 points on SAT mathematics scores reported between 1970 and 1976, and a 35 point difference in their own study. The data indicate that about 60% of the gap in quantitative performance is due to course differences in high school and the authors link this to differential sex-role socialization of boys and girls at earlier ages. The question here is whether the different socialization, the likely difference in coursework, and the residual male-female discrepancy in mathematics scores unexplained by coursework differences might cause male and female mathematics and science teachers to rate the importance of science and mathematics tasks differently on a job analysis.

Second, the response by race should be evaluated for differences because TCTs are usually challenged in court as a result of adverse minority impact. In LULAC v. State of Texas (1985), a preliminary injunction (later overturned) was granted against the use of a basic skills test for undergraduates seeking to enroll in teacher education courses that adversely impacted minority students. One criticism that plaintiffs had regarding the validation study was that the survey responses to questions about adequacy of preparation for the test had not been broken down by race. When there is adverse impact, demonstration of the representativeness of the sample will become an important issue if there are suspected differences in job analysis responses by race, region, gender, or some other classification considered arbitrary under Title VII or the 14th Amendment.

Third, a recent EEOC determination (EEOC, 1988) stated that the Texas education agencies charged in the complaint:

> have discriminated against Blacks and persons over 40 years of age who took the TECAT [Texas Examination of Current Administrators and Teachers] in 1986 and were removed from their teaching positions as a result. Accordingly, we find that Title VII and the ADEA have been violated as to Charging party and all similarly situated individuals. (p. 2)

The ADEA (Age Discrimination in Employment Act of 1967) prohibits age discrimination in employment. Although amendments to the ADEA have raised and finally eliminated a maximum age for which protection applies, 40 has remained the minimum age at which protection from age discrimination begins (Fretz and Dudovitz, 1987). Response differences associated with age of rater may contribute to bias in content domain definition.

In summary, there is little guidance in the literature regarding the adequacy or quality of a job analysis. As job analysis has become more important to the validity argument of an increasing number of high-stakes tests, several questions have been raised in the literature regarding job analysis methodology, adequacy, and fairness. Questions have also been raised regarding the relationship of person-characteristics, including race, to job analysis responses. In the legal arena, where issues of test bias and adverse impact are discussed, the validity of the job analysis responses has become an issue in test-related court cases (Kuehn, et al. 1989).

Analysis Samples

The Georgia Teacher Certification Tests (TCT) are undergoing revision. As a first step in this process, the Georgia Job Analysis Questionnaire (JAQ) was distributed by Georgia Assessment Project (GAP) in January, 1987, to all certified personnel identified by the Georgia Department of Education. Questionnaires were sent to 2801 math and 2468 science teachers or teacher supervisors identified as certified in these fields. Twenty-five different questionnaire forms had been developed to represent the fields of certification in the state. On each form, the first 54 task statements were the same and included teacher activities identified as common to the profession. The remainder of the task statements were related to the content of each field. The task statements were written by GAP staff with the help of content experts and are based on a review of the literature in the field, the school curricula, and classroom observations. The science form had a total of 148 task statements and the math form had 160 statements. In addition to the task statements, the questionnaires included places to code biographical data. For science and math respectively, 1384 and 1600 usable responses were available. Table 1 summarizes the characteristics of both groups of respondents.

On the Georgia JAQ, teachers rated the task statements by indicating for each one whether they actually performed the task, how important it was to the learning process on a one to five scale (little, some, moderate, considerable, or great importance), and whether a minimally competent teacher should be able to perform the task throughout job tenure.

Task Statement Content Clusters

Analyses were done on average ratings of clusters of related

task statements. Science task statements were grouped into five broad categories by science teachers and teacher educators at a meeting held to begin revision of the current TCT. In the sorting done by the science content experts, task statements could fall into more than one category because of the interrelated nature of the science fields. For the purposes of this study, The 15 task statements that overlapped categories were eliminated and the remaining 79 that the content experts generally agreed represented one content area were retained. These fell into the following five content clusters: general scientific processes (17 tasks), biology (21 tasks), chemistry (14 tasks), physics (16 tasks), and earth science (11 tasks). These generally represent the curriculum in grades 7 through 12, the range covered by the secondary certificates.

Revision meetings have not yet been held in secondary math so the categories indicated by two math teachers who worked in the development of the task statements were used to cluster math tasks and eliminate those that overlapped content categories. These teachers judged 80 of the tasks to fall into one of six categories: basic math concepts (21), algebra (19), geometry (10), trigonometry (8), calculus (5), and computers and programming (17). The remaining 26 task statements were not used in the analysis of the mean responses for the content clusters. The mean importance ratings for each content area are plotted in Figures 1 through 6.

Multivariate analysis of variance profile analysis was used to evaluate rating differences of the mean importance of the science and math task clusters by the three independent variable groupings (Harris, 1985, and Nunnally, 1978). This allowed for tests of significance for group by content cluster interactions (parallelism test), group mean differences in importance ratings (main effects or levels test), and differences in the ratings of the content clusters (flatness test). The levels test, or test of group mean differences, is a univariate test and is reported as such. The other two tests in profile analysis (parallelism and flatness) are multivariate tests and the results are reported accordingly as Wilks λ , its associated F, and its significance. Where significant main effects were found for groups, analysis of variance was used to discover the simple effects of group differences at each content cluster rating. Results of the MANOVAs and ANOVAs are reported in Tables 2 through 12 for science and math. These tables appear under the appropriate plot of mean responses for each variable. Table 13 summarizes the significant multivariate results found in the analyses in both fields.

Because of the likelihood of finding significant results with the large sample sizes in this study, estimates of effect size were calculated where any significant result was found. The general formula suggested by Maxwell, Camp, and Arvey (1981) for estimating strength of association (omega squared) in factorial designs is:

$$\omega^2 = \frac{\text{SS effect} - (\text{df effect} \times \text{MSw})}{\text{SS total} + \text{MSw}}$$

For the calculation of ω^2, the averaged univariate tests sums of squares were used for the parallelism and flatness tests. Table 14 summarizes the estimated effect size or variability accounted for for the significant results.

In his discussion of effect size, Cohen (1977) states that it is not common in the behavioral sciences to see an effect size as large as .25. The frame of reference he suggests is that an effect size of .01 is considered small, .06 is medium, and .16 or greater is considered a large effect size in behavioral research. The results of this study are interpreted according to this standard.

Interaction of Importance Ratings by Group

As summarized in Table 13, the test of profile parallelism or interaction showed significant results (nonparallel profiles) for all variables except age in science. However, as shown in Table 14, the significant interactions did not have effect sizes that reached what Cohen defines as the small range.

Main Effects: Group Mean Differences

In profile analysis, where significant interactions (lack of parallelism) are found, the main effects (group differences) or flatness tests (dependent variable or content rating differences) are difficult to interpret and may not be meaningful (Harris, 1985 and Nunnally, 1978). However, because of the exploratory nature of this work, both are evaluated here. In addition, Keppel (1982) states that:

> If there is no interaction, or if the interaction is significant but trivial [the case with these results], the outcome of the F tests involving the main effects can be interpreted without qualification. With a sizable and significant interaction, on the other hand, the meanings of these F tests must be interpreted with caution. (p. 211)

Significant main effects or group (sex, race, age) differences were found for sex and race in science but for none of the group differences in math (Table 13). Table 14 shows that only the race difference (Black respondents rated tasks higher in importance) in science had a meaningful effect size. Other researchers have reported similar results. Veres (1983) found that where there were racial differences in ratings, Blacks also rated the tasks of higher importance. Rosenfeld et al. (1986) report mean ratings of pedagogical tasks by race (Appendix J, Table 4) but did not test differences for significance. On all variables, Black teachers rated the tasks of higher importance

than did White teachers. Elliot (1987) did not report importance ratings and did not report analyses of the time-spent ratings by incumbent characteristics such as race, sex, or age.

In both fields, significant differences in mean content cluster ratings were found. There is a definite hierarchy of importance of the content of each field. In both fields, the basic or core content is rated the highest in importance. The content taught with less frequency is rated lower in importance. These response differences could be due to a perception of importance to student learning based on the small number of students who actually study these higher level subjects or the differences could be due to the teachers' familiarity with the content. If teachers are less knowledgeable about the higher level course content, they may be rating it lower in importance for that reason. As shown in Table 14, the differences in science content rating show a small effect size ($\omega^2 = .03$) while the math content ratings show a moderate effect size ($\omega^2 = .14$).

Simple Effects

To explore the source of group rating differences in each content area, tests of simple effects of each group were conducted using oneway ANOVA where the profiles were nonparallel and the overall group means were significantly different. To avoid the proliferation of Type I error that can be the result of multiple tests, the decision was made to suspend judgment on the significance of Fs that fell between .05 and an adjusted alpha level (Keppel, 1982, p. 163). Using the Bonferroni adjustment (Harris, 1985, p. 8), the critical alpha is set at .05 divided by the number of univariate tests conducted, or .01 for science and .0083 for math. Significance is indicated by probabilities smaller that these adjusted levels, judgment is suspended for probabilities in the corrected alpha to the .05 range, and nonsignificance is concluded for probabilities greater than .05. The results are reported in this manner in the tables associated with these tests.

Inspection of the results of these tests of simple effects for science (Tables 3 and 5) does not yield any generalizations about the content areas in which differences in importance ratings lie. In other words, the likelihood of any particular content area being rated differently by one group or another does not seem to be related to the frequency with which that content is taught. For race in science, the only variable with a meaningful effect size, all the simple effects were significant except for the earth science rating.

In math, the simple effects tests are reported in Tables 8, 10, and 12. All the math interactions were significant. Females rated the more basic math courses higher than males. The ratings were not different for the higher level math courses. For race, there is no clear pattern of differences and for the age categories, older teachers rated only geometry higher in importance.

Conclusions

The results show that the content areas rated by teachers on the questionnaire vary in their relative importance either because of teacher familiarity with the content or because of the frequency with which the content is taught by the teachers.

Regarding group differences and interactions between the groups and the content ratings, significant results were found for sex and race in science. However, when effect size was estimated using ω^2, meaningful results were found only for race in the science ratings. No important sex rating differences were found although review of the literature showed that there is evidence of differential mastery of science and math content and differential test performance for males and females in these areas. There are no important age effects in the ratings.

Because of the potential for adverse impact in selection procedures, racial differences in the ratings are of major concern. The test content domain, as defined by the job analysis responses, appears not to be biased. Of particular importance is the demonstration that Black job incumbents on average do not rate job content task statements lower in importance than White job incumbents. Content domain definition based on importance ratings would not yield test content considered unimportant by minority test-takers. The validity of the content domain and of the job analysis is supported by the absence of important group response differences.

References

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

American Psychological Association, Division of Industrial-Organizational Psychology (1987). Principles for the validation and use of personnel selection procedures (3rd ed.). College Park, MD: Author.

Barrett, R.S. (1980). Is the test content-valid: Or, does it really measure a construct? Employee Relations Law Journal, 6(3), 459-475.

Bemis, S. E., Belenky, A. H., & Soder, D. A. (1983). Job analysis. Washington, DC: Bureau of National Affairs.

Boyles, W. R., Palmer, C. I., & Veres, J. G. (1980, July). Bias in content-valid tests. Paper presented at the International Personnel Management Association Assessment Council Annual Conference, Boston.

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

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cross, L. H. (1985). Validation of the NTE tests for certification decisions. Educational Measurement: Issues and Practice, 4(3), 7-10.

Elliot, S. M. (1987). Validating job analysis survey instruments used in developing teacher certification tests: A construct validity study. Paper presented at the National Council of Measurement in Education Annual Conference, Washington, DC.

Equal Employment Opportunity Commission (1988). Letter of Determination for Bradley against Tyler Independent School District et al. Dallas: District Office, EEOC.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice (1978). Uniform guidelines on employee selection procedures. Federal Register, August 25, 43(166), 38290-38315.

Fretz, B.D. & Dudovitz, N.S. (1987). The Law of Age Discrimination. Chicago: National Clearinghouse for Legal Services, Inc.

Guion, R. M. (1978). Scoring of content domain samples: The problem of fairness. Journal of Applied Psychology, 63(4), 499-506.

Harris, R. J. (1985). A primer of multivariate statistics (2nd ed.). Orlando: Academic Press.

Kane, M.T., Kingsbury, C. Colton, D. & Estes, C. (1989). Combining data on criticality and frequency in developing plans for licensure and certification examinations. Journal of Educational Measurement, 26(1), 17-27.

Keppel, G. (1982). Design and analysis: A researcher's handbook (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Kuehn, P.A., Stallings, W.M. & Holland, C.L. (1989). Court-defined job analysis standards in content validation. Paper presented at NCME Annual Meeting, March 28, 1989, San Francisco.

Lehmann, I. J., & Phillips, S. E. (1987). A survey of state teacher-competency examination programs. Educational Measurement, 6(1), 14-18.

LULAC v. State of Texas, 628 F. Supp. 304 (1985).

Madaus, G. (1987). Teacher certification tests: Do they really measure what we need to know? Phi Delta Kappan, 69(1), 31-38.

Maxwell, S. E., Camp, C., & Arvey, R. D. (1981). Measures of strength of association: A comparative examination. Journal of Applied Psychology, 66(5), 525-534.

Mehrens, W. A. (1986). Validity issues in teacher competency tests. Unpublished manuscript. Michigan State University.

Mehrens, W. A. (1987). Validity issues in teacher licensure tests. Journal of Personnel Evaluation in Education, 1(2), 195-229.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Pallas, A. M., & Alexander, K. L. (1983). Sex differences in quantitative SAT performance: New evidence on the differential coursework hypothesis. American Educational Research Journal, 20(2), 165-182.

Potter, D. (1980). Job analysis of teaching. Final Report. Princeton, NJ: Educational Testing Service.

Prien, E. P. (1977). The function of job analysis in content validation. Personnel Psychology, 30, 167-174.

Rosenfeld, M., Thornton, R. F., and Skurnik, L. S. (March, 1986). Analysis of the professional function of teachers: Relationships between job functions and the NTE Core Battery. Research Report 86-8. Princeton, NJ: Educational Testing Service.

Tenopyr, M. L. (1986). Needed directions for measurement in work settings. In B. S. Plake & J. C. Witt (Eds.), The future of testing (pp. 269-288). Hillsadale, NJ: Erlbaum.

Veres, J. G., III, Boyles, W. R., & Champion, C. H. (1983, May). Bias in content-valid tests revisited. Paper presented at the International Personnel Management Association Assessment Council Annual Conference, Washington, DC.

Veres, J. G., III, (1983). Bias in content validity: An investigation into sources of racial differences in job analysis (Doctoral dissertation, Auburn University, 1983). Dissertation Abstracts International, 44, (3-B) 946.

Table 1

Gender, Race, and Age Breakdown for Questionnaire Respondents in Percentages

	Science	Math
Gender		
Male	39%	27%
Female	61	73
Race		
Black	23	18
White	77	82
Age		
$\leq$ 40	61	64
$>$ 40	39	36

Figure 1. Science: Mean task ratings by sex.

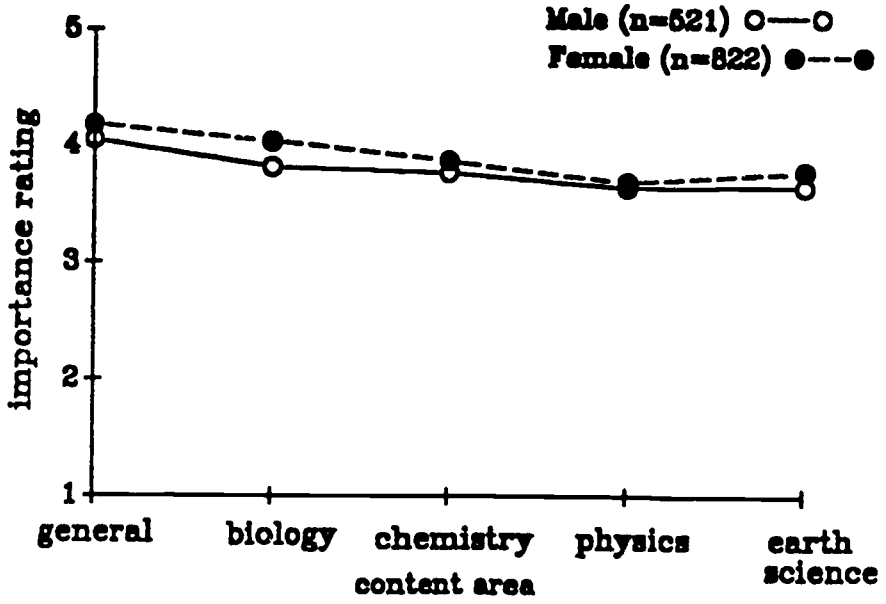

Table 2

Profile Analysis for Science Content by Sex

Test	Wilks λ	MS	MS error	F	df	Sig.
Parallelism	.991			2.987	4,1275	.018
Flatness	.722			122.702	4,1275	.000
Levels *		4.743	.579	8.187	1,1278	.004

* univariate test

Table 3

Univariate Tests for Simple Effects of Sex

Content Area	F	df	Sig.	Decision
General	11.100	1	.001	**
Biology	22.258	1	.000	**
Chemistry	3.269	1	.071	NS
Physics	.735	1	.391	NS
Earth Science	6.206	1	.013	*

NS not significant
* suspend judgment
** significant

Figure 2. Science: Mean task ratings by race.

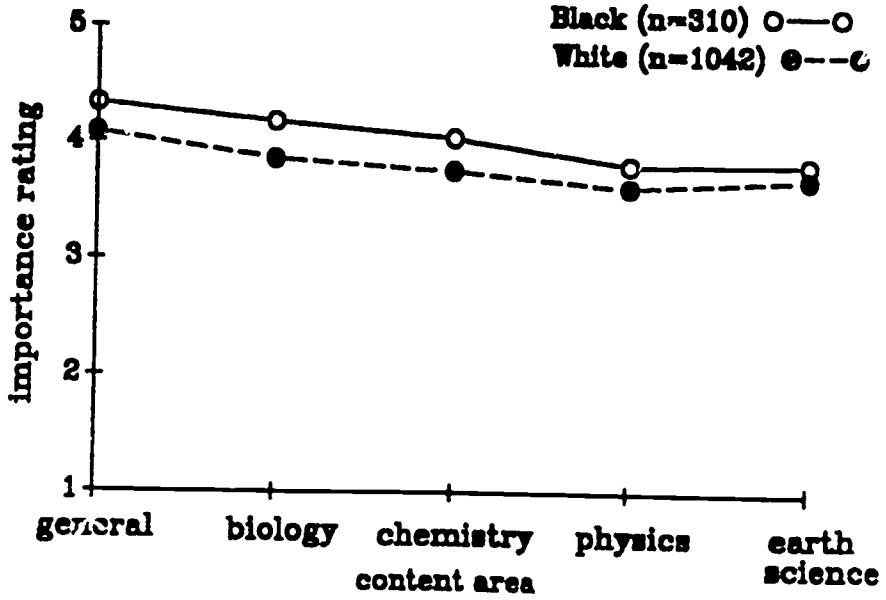

Table 4

Profile Analysis for Science Content by Race

Test	Wilks λ	MS	MS error	F	df	Sig.
Parallelism	.988			3.782	4,1286	.005
Flatness	.719			125.442	4,1286	.000
Levels *		11.532	.577	19.973	1,1289	.000

* univariate test

Table 5

Univariate Tests for Simple Effects of Race

Content Area	F	df	Sig.	Decision
General	32.286	1	.000	**
Biology	29.508	1	.000	**
Chemistry	21.218	1	.000	**
Physics	7.602	1	.006	**
Earth Science	3.438	1	.064	NS

NS not significant
* suspend judgment
** significant

Figure 3. Science: Mean task ratins by age.

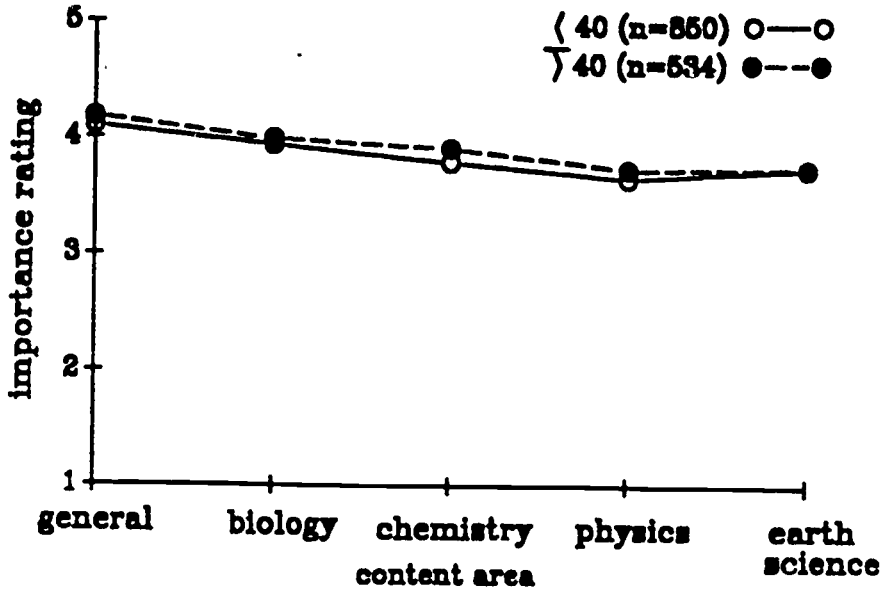

Table 6

Profile Analysis for Science Content by Age Category

Test	Wilks λ	MS	MS error	F	df	Sig.
Parallelism	.994			2.058	4,1315	.084
Flatness	.715			130.760	4,1315	.000
Levels *		1.258	.585	2.150	1,1318	.143

* univariate test

Figure 4. Mathematics: Mean task ratings by sex.

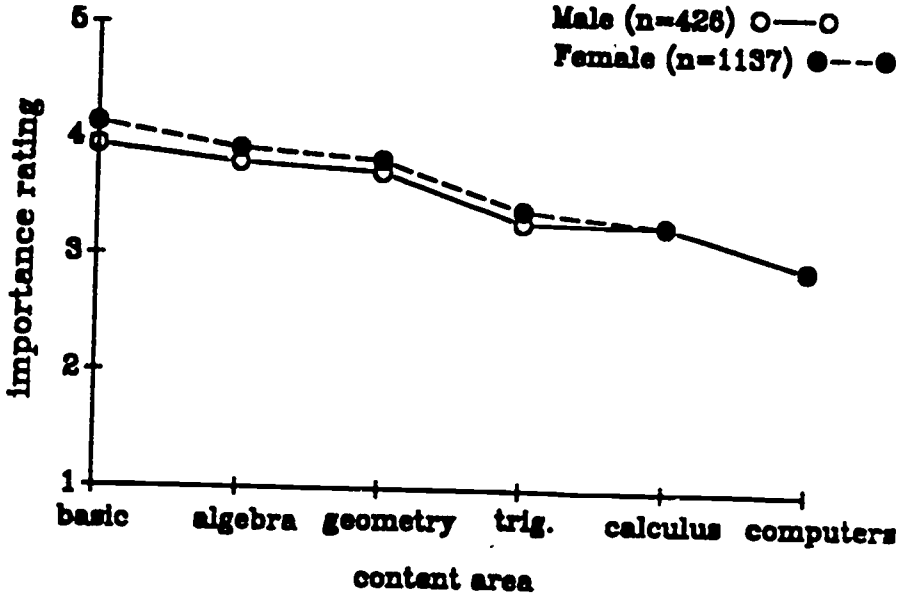

Table 7

Profile Analysis for Math Content by Sex

Test	Wilks λ	MS	MS error	F	df	Sig.
Parallelism	.981			5.523	5,1417	.000
Flatness	.363			496.751	5,1417	.000
Levels *		2.043	.647	3.159	1,1421	.076

* univariate test

Table 8

Univariate Tests for Simple Effects of Sex

Content Area	F	df	Sig.	Decision
Basic	30.561	1	.000	**
Algebra	6.574	1	.010	*
Geometry	4.416	1	.036	*
Trigonometry	2.556	1	.110	NS
Calculus	.041	1	.840	NS
Computers	.001	1	.975	NS

NS not significant
* suspend judgment
** significant

Figure 5. Mathematics: Mean task ratings by race.

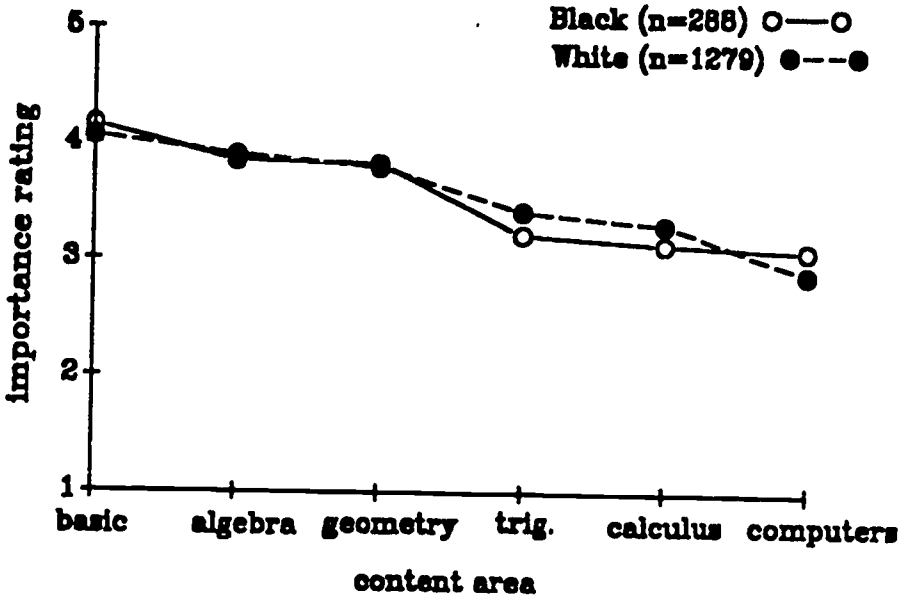

Table 9

Profile Analysis for Math Content by Race

Test	Wilks λ	MS	MS error	F	df	Sig.
Parallelism	.972			8.158	5,1424	.000
Flatness	.366			494.368	5,1424	.000
Levels *		.030	.650	.046	1,1428	.830

* univariate test

Table 10

Univariate Tests for Simple Effects of Race

Content Area	F	df	Sig.	Decision
Basic	7.156	1	.008	**
Algebra	.992	1	.319	NS
Geometry	.418	1	.518	NS
Trigonometry	6.819	1	.009	*
Calculus	4.065	1	.044	*
Computers	7.399	1	.007	**

NS not significant
* suspend judgment
** significant

Figure 6. Mathematics: Mean task ratins by age.

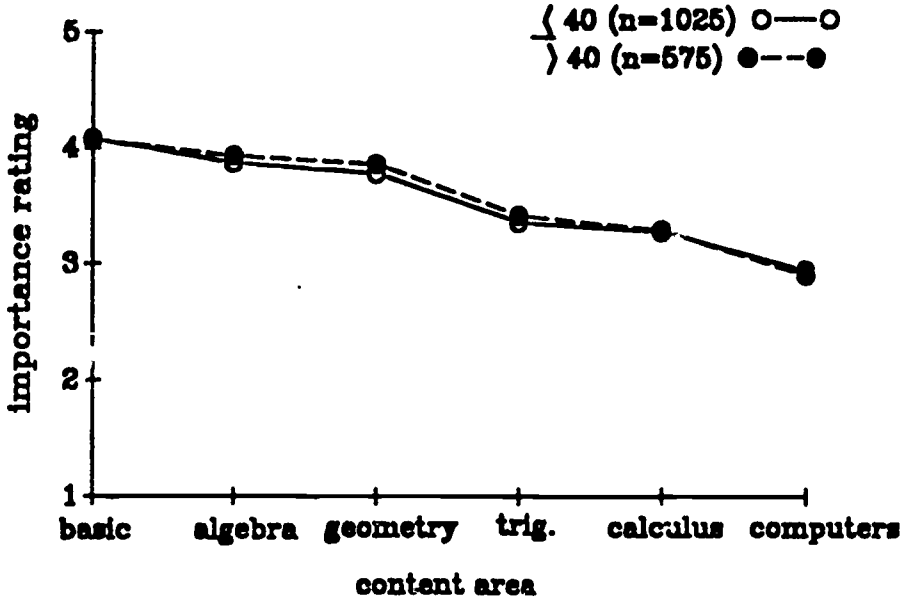

Table 11

Profile Analysis for Math Content by Age Category

Test	Wilks λ	MS	MS error	F	df	Sig.
Parallelism	.988			3.431	5,1449	.004
Flatness	.367			499.193	5,1449	.000
Levels *		.180	.649	.278	1,1453	.598

* univariate test

Table 12

Univariate Tests for Simple Effects of Age Category

Content Area	F	df	Sig.	Decision
Basic	.137	1	.712	NS
Algebra	1.863	1	.172	NS
Geometry	4.270	1	.039	*
Trigonometry	.733	1	.392	NS
Calculus	.018	1	.895	NS
Computers	.424	1	.515	NS

NS not significant
* suspend judgment
** significant

Table 13

Summary of Significant Profile Analysis Results

	Interaction	Group Difference	Subarea Difference
		Science	
Incumbent Characteristics			
Sex	*	*	*
Race	*	*	*
Age			*
		Math	
Sex	*		*
Race	*		*
Age	*		*

* $p < .05$

Table 14

Estimates of Variability Accounted For (ω^2) for Significant Results

		Interaction	Group	Content Area
Incumbent Characteristics				
Sex	Science	.00067	.00399	.0329 *
	Math	.00077	-	.1432 **
Race	Science	.00064	.0103 *	.0333 *
	Math	.00278	-	.1437 **
Age	Science	-	-	.0342 *
	Math	.00021	-	.1434 **

* small effect size
** medium effect size
Cohen (1977)