Developing g-loaded selection tests without adverse impact.

Michael D. Biderman

Bart Weathington

University of Tennessee at Chattanooga

Authors' Note: Correspondence regarding this article should be sent to Michael Biderman, Department of Psychology / 2803, U.T. Chattanooga, 615 McCallie Ave., Chattanooga, TN 37403. Tel.: (423) 425-4268. Email: Michael-Biderman@utc.edu

Paper submitted for presentation at the 25th Annual Conference of The Society for Industrial and Organizational Psychology, Atlanta, GA 2010.

Poster

TITLE

Developing g-loaded selection tests without adverse impact

ABSTRACT

Individual job knowledge items that were both valid and that had little adverse impact were used to develop selection tests. The tests were nearly as valid as those chosen based on validity and had considerably reduced adverse impact – non significant in two separate banks of questions.

PRESS PARAGRAPH

Many organizations avoid the use of job knowledge and other types of selection tests with high loading on general cognitive ability (g) because use of the tests often result in adverse impact on minority applicants. Using two large item banks, we selected items that were both valid and had little adverse impact. Tests based on the specially selected items were nearly as valid as those based on validity alone and in contrast to the validity based tests, had nonsignificant adverse impact.

While it is commonly accepted that cognitive ability tests are valid predictors of job performance, there is significant disagreement about how and when these tests should be used (Murphy, Cronin, & Tam, 2003). The major disadvantage to the use of mental ability measures is that adverse impact is often found when these evaluations are used to make employment decisions (Hough, Oswald, & Ployhart, 2001; Roth, Bevier, Bobko, Switzer, & Tyler, 2001). In the past many attempts have been made to address adverse impact issues by using approaches such as race norming and banding. However, the Civil Rights Act of 1991 made race norming illegal and banding has its own issues with administration (see Bobko, Roth, & Nicewander, 2005). Identified race differences in cognitive ability results, however, depend to a large extent on the complexity of the job (Hough et al.)

It has been argued that jobs are becoming less complex and correspondingly cognitive ability is becoming less important for predicting future work performance. However, analyses suggest that this is only occurring in some situations and other jobs are actually becoming more complex and cognitively demanding (Landy, 2007). For this reason it is especially critical to evaluate the use of cognitive ability and job knowledge tests for jobs such mechanics and electricians that involve significant complexity, deductive reasoning, and problem solving ability.

When dealing with complex jobs, and with a sufficiently large pool of items, it should be possible to identify items that maximize the prediction of on-the-job performance while at the same time minimizing race based differences in item responses. Accordingly, this paper documents the creation and validation of job knowledge tests for mechanics and electricians at a large manufacturing company. The goal of the research was to develop job knowledge tests that were valid with minimal adverse impact.

METHOD

Participants.

Participants were 172 electricians and 235 mechanics in a large manufacturing facility in the Southeastern United States. Mechanics' jobs required maintenance of the machinery involved in the manufacturing processes. Electricians maintained electrical aspects of the machinery. Of electricians, 83% were White, 11% Black, 3% Hispanic and 3% missing or listed as Other. Of mechanics, 73% were White, 19% Black, 6% Hispanic and 2% missing or listed as other.

Test items.

A pool of 2,000 thousand test items for each job type served as the populations of items from which items used in tests were drawn. From the initial pool, 1101 items were drawn for possible use in mechanics' tests and 1560 were drawn for possible use in electricians' tests. The initial selection of items was conducted by a person acquainted with jobs of both mechanics and electricians.

Subject matter experts (SMEs) evaluated the appropriateness of items in the initial sample for use in tests for mechanics and electricians. The SMEs for electrician items were employees who supervised electricians. Those for mechanics items were supervisors of mechanics. SMEs placed each item into one of three categories, with the first category containing items the SME felt were appropriate for the test, the second containing items the SME felt might be appropriate, and the third containing items the SME felt would not be appropriate. At least three SMEs rated each item. The mean category rating for each item was computed and the items were rank-ordered based on SME category mean ratings. Items were picked for inclusion in the tests from the top of the rank-ordered list.

A sample of 913 items was drawn from the rank ordered list of items for the tests to be used for electricians. A similar sample of 676 items was drawn for mechanics. The item samples had different sizes because tests created for the organization were to be for three levels of mechanics and four levels of electricians with about 200 items for each level. For this research, however, the differences between levels were ignored.

All the test items rated highest by SMEs were administered to electricians and mechanics working the organization. Each electrician answered 913 items and each mechanic answered 676 items. Testing was spread out over a whole day. Each item was scored as either incorrect with value 0 or correct with value 1.

Performance Evaluations.

The overall performance of each of the employees participating in the validation was evaluated by two persons familiar with both that employee's performance and the performance of other employees. Criterion scores were obtained by having evaluators first categorize each employee as either "Among the best", "Above average", "Average", "Below average", or "Among the weakest". After that categorization, each employee was assigned a number representing his/her position with the category. Larger numbers were assigned to better employees. The numbers 41-50 were used for those employees categorized as "Among the best", 31-40 for those categorized as "Above average", 21-30 for those categorized as "Average", 11-20 for those who were "Below average", and 1-10 for those who were categorized as "Among the weakest".

For electricians, the correlation of the two sets of evaluations was .625. For mechanics, the correlation was .819. Given the high positive correlations between the pairs of evaluations, an overall evaluation was computed as the mean of the two, resulting in one criterion score for mechanics and one criterion score for electricians. Figure 1 presents distributions of criterion scores for each the group.

Item Validity

The correlation between score on an item and the performance evaluations was computed for each item. Thus 913 correlations were computed for electrician items and 676 correlations for mechanics items. Those correlations were the item validities in the analyses that follow. Figure 2 presents distributions of the item validities for both job types.

Tests Based only on Validity.

Each set of items was ordered by validity and within each set items were drawn top down from the ordered lists to create a 50-item test for electricians and a 50-item test for mechanics. The tests created only from the most valid items will be called the Validity Only tests.

Item Adverse Impact.

An index of adverse impact was created for each item. For the purpose of this index, employees whose ethnic identification was listed as White were so categorized, while those whose ethnic identification was either Black or Hispanic were categorized as Nonwhite. Those who listed some other ethnic identification or none were excluded from the analyses involving this index. The index was simply the correlation of item correctness with the dichotomous White vs. Nonwhite dichotomy. The adverse impact of an item was thus defined as its correlation with the ethnic dichotomy. For that dichotomy, Whites were scored as 1 and Nonwhites as 0, resulting in items favoring Whites having positive adverse impact scores. Figure 3 presents distributions of adverse impact values of the items for both job types.

Tests Equated for Adverse Impact.

To create tests with minimal adverse impact, called Low AI tests from now on, each set of items was ordered in descending order by validity and in ascending order by the absolute value of adverse impact. From that ordered list, items were selected from the top down with the restriction that signed adverse impact had to be less than a criterion value. That value was + .01 for electricians and +.05 for mechanics items. The first 50 items from each list were included in the Low AI test for that list. Scatterplots of individual item validity vs. individual item adverse impact are presented in Figure 4. Filled circles in Figure 4 represent the selected items for each Low AI test.

RESULTS

Reliabilities and validities of the Validity Only tests and Low AI tests are presented in Table 1. As shown in the table, the Validity Only tests are acceptably reliable. Reliability of the Low AI test for electricians was marginally acceptable while the reliability of the low AI test for mechanics certainly could be improved.

Validities of the two tests are also presented in Table 1. Interestingly, validities of the low adverse impact tests were essentially equal to validities of the tests based on validities. In fact, the validity of the low AI test for mechanics was slightly higher than that of the test created without consideration of AI.

Inspection of Figure 4 indicates that adverse impact is positively correlated with validity for both collections of items. Those correlations are .33 and .37 for electricians and mechanics items respectively.. The positive correlations between adverse impact and validity mean that selection based on validity only would result in a biased sample of items with respect to adverse impact, with a majority of the items having positive adverse impact. This bias in items is reflected in overall test scores obtained by Whites and Nonwhites in both groups. Means of Whites and Nonwhites on the two tests are presented in Table 2. Most importantly for the present study are the comparisons of adverse impact between the Validity Only and Low AI tests. For both electricians and mechanics, the standardized difference between White and Minority means was very large and statistically significant on the Validity Only tests, with standardized d values of .80 and .74 for electricians and mechanics. The difference, however, was considerably smaller for the Low AI tests, with d being reduced to .21 for the electricians test and to .11 for the mechanics test. For both tests, the difference between White and Minority means was not significant. Based on the sample sizes available, power to detect a medium effect size was .70 for electricians and .91 for mechanics.

DISCUSSION

This initial exploration of the development of job knowledge tests that are both valid and without adverse impact suggests that it may be possible to “have ones cake and eat it too”. This is especially relevant for jobs of a complex nature. Job knowledge tests are particularly important for jobs such mechanics and electricians. However, the relationship of these tests with cognitive ability and related adverse impact considerations has the potential to limit the use of otherwise valid tests. Selecting items for low adverse impact in the manner shown here may expand the use of such tests.

The validities of the 50-item tests presented in Table 1 are slightly smaller than those presented by Schmidt and Hunter (1998) for job knowledge tests used alone. The value presented by Schmidt and Hunter was .48, compared with the range of .42 to .46 for the tests developed here. We note that the Schmidt and Hunter values were corrected for downward bias due to measurement error and range restriction. Since the job performance measure was based on only two evaluations, we felt it inappropriate to correct the validities presented in Table 1.

However, if reliability of the performance measure for each job type were .80, for example, the validities in Table 1 would be corrected to range from .47 to .51, values quite comparable with those reported by Schmidt and Hunter (1998).

The distributions of adverse impact shown in Figure 3 suggest that if test developers have access to large enough item pools, it is certainly advantageous to search for items that have low adverse impact or, cautiously to balance out other items, even reverse adverse impact. The existence of individual differences in AI raises the psychometric question of the generalizability of such item differences across groups. Based on the now well accepted principle that validity generalizes across different respondent groups, we would expect that adverse impact would also exhibit generalizability. That would certainly be a prerequisite for use of the results presented here for long term test use of a test. Unfortunately, our short window of opportunity with the participating organization prevented us from examining the extent to which absence of adverse impact of the Low AI tests created here generalized to other electricians and mechanics. Such an investigation would be an important part of future explorations of the possibility of low AI tests.

The identification of item differences in adverse impact creates the possibility of identification of the sources of group differences. Thus another avenue of future research would be to examine item characteristics, such as reading level, wording, and specific content to identify item correlates that determine level of AI. Identifying such characteristics would make it easier to reduce adverse impact as done here and also might shed light on the nature of differences between groups that result in adverse impact.

As Hough et al. (2001) point out, many factors such as measurement method, cultural issues, and stereotypes contribute to group differences in performance on cognitive ability related measures. However, this relationship is at least partially influenced by job complexity. As our results demonstrate, in some situations tests for complex jobs that are both valid and have minimal adverse impact can be developed based on g-loaded items.

REFERENCES

Bobko, P., Roth, P. L., & Nicewander, A. (2005). Banding selection scores in human resource management decisions: Current inaccuracies and the effect of conditional standard errors. *Organizational Research Methods, 8,* 259-273.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, dectection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *International Journal of Selection and Assessment, 9,*152-194.

Landy, F. L. (2007). The validation of personnel decision in the twenty-first centure: back to the future. In S. M. McPhail (Ed.), *Alternative validation strategies* (pp.409-426). San Francisco: Jossey-Bass.

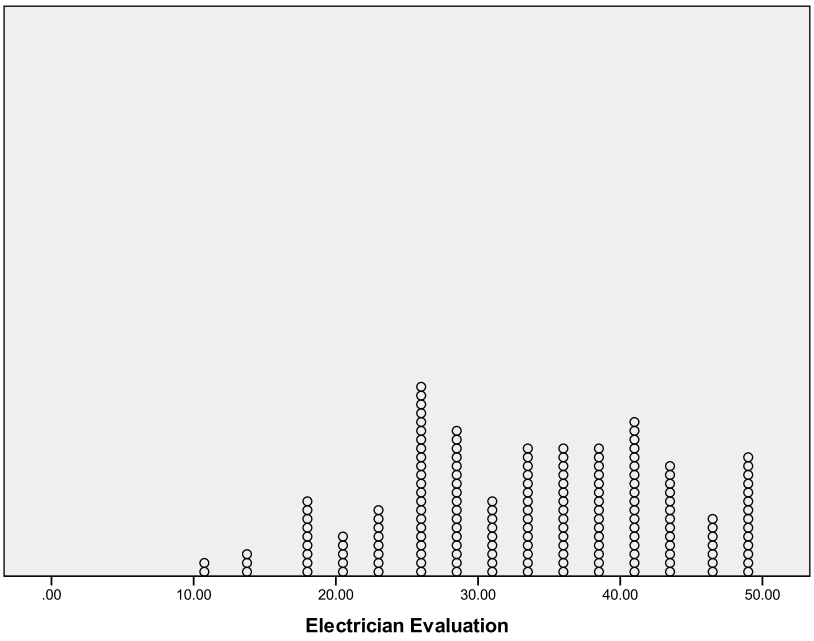
Murphy, M. R., Cronin, B. E., & Tam, A. P. (2003). Controversy and consensus regarding the use of cognitive ability testing in organizations. *Journal of Applied Psychology, 88,* 660-671.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54,* 297-330.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

Figure 1. Distributions of performance evaluation scores for electricians and mechanics.

Electricians



Mechanics

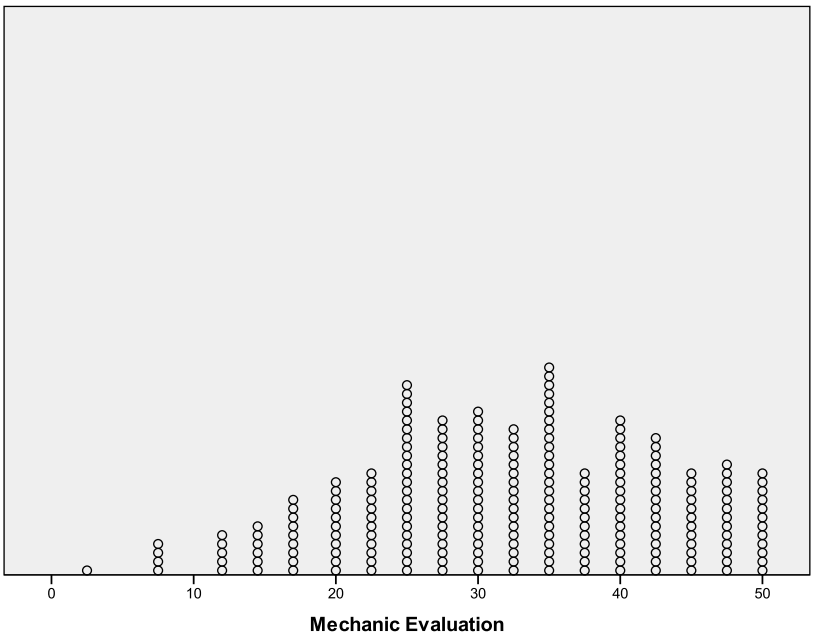
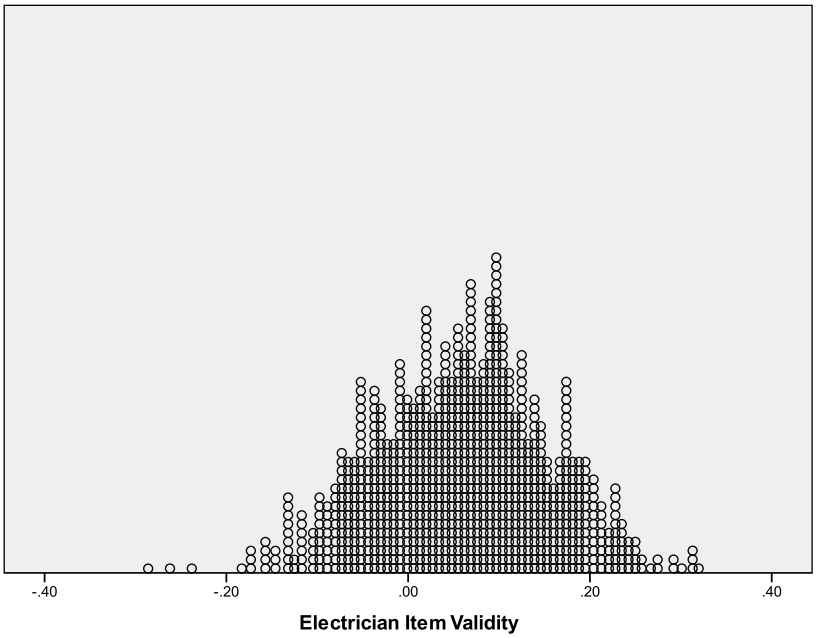


Figure 2. Distributions of individual item validities for electricians and mechanics items.

Electricians



Mechanics

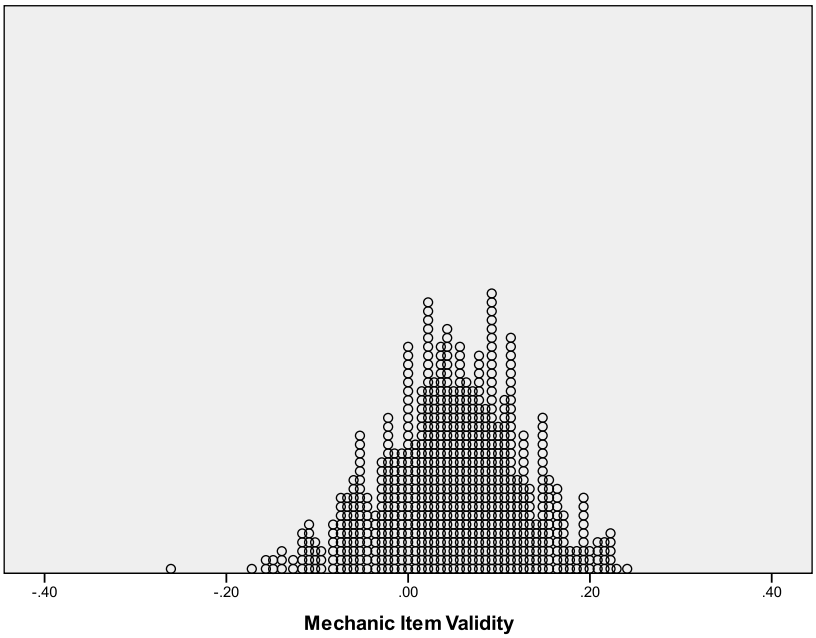
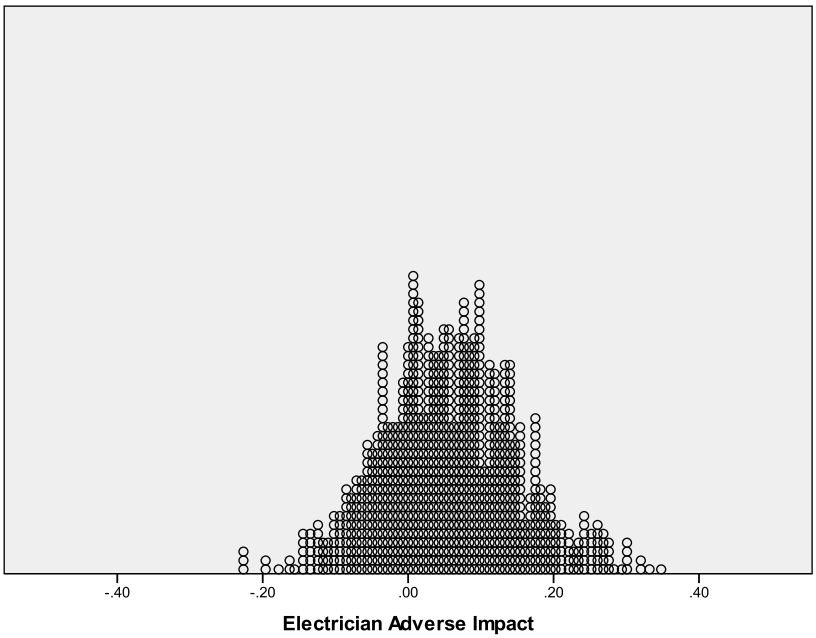


Figure 3. Distribution of individual item adverse impact values for electricians and mechanics items.

Electricians



Mechanics

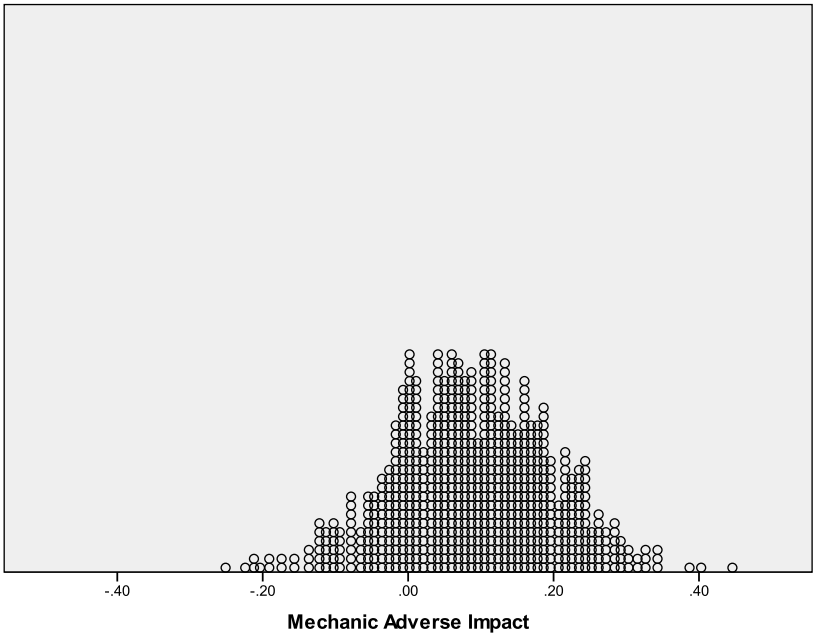


Figure 4. Scatterplot of individual adverse impact vs. validity for electrician and mechanics items. Filled circles represent items used for low adverse impact test.

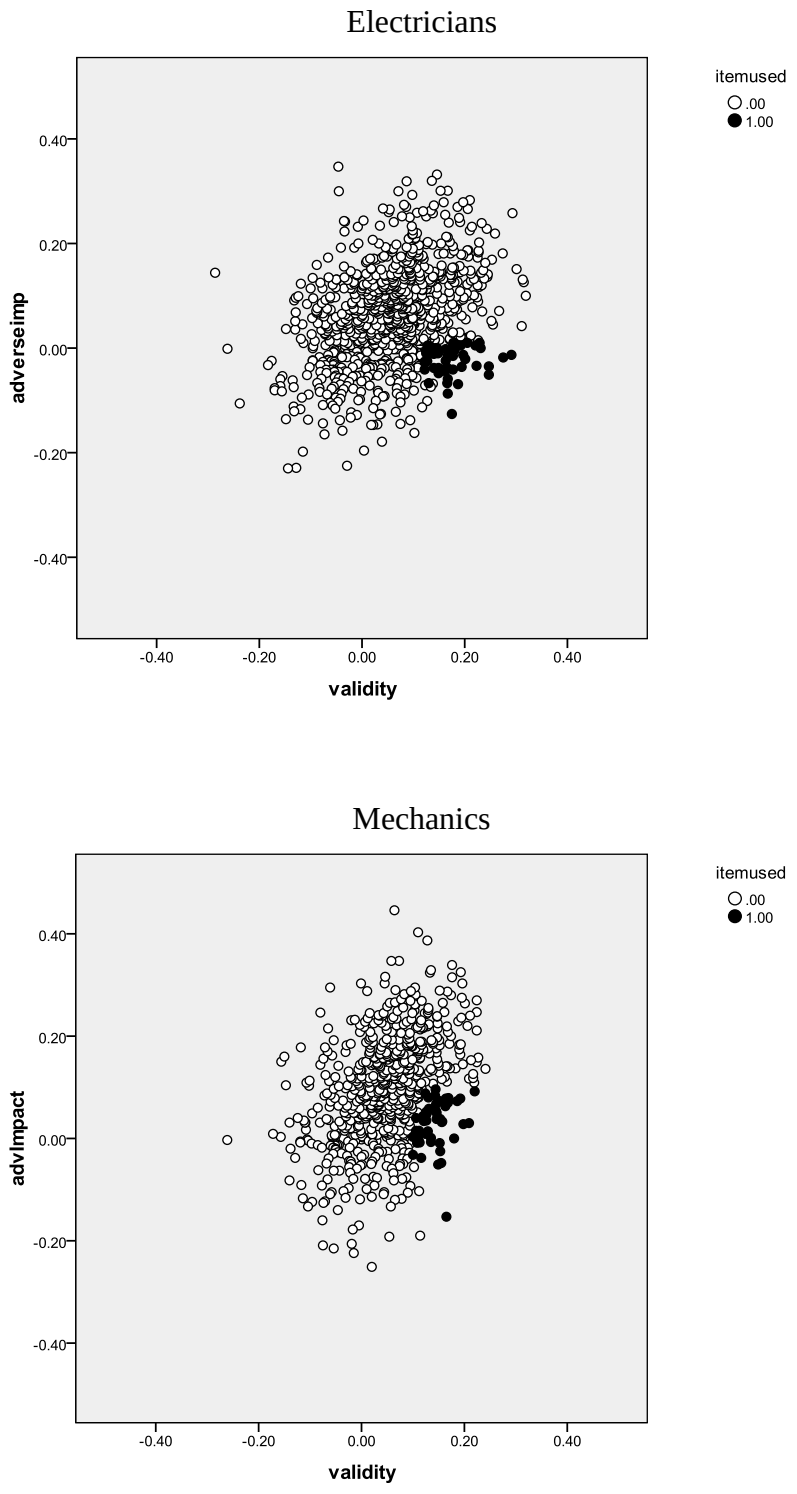


Table 1. Validities and reliabilities of 50-item tests. Values corrected

	Validity Only	Low AI
Electricians		
Reliability	.88	.78
Validity	.44	.42
Mechanics		
Reliability	.88	.67
Validity	.42	.46

Table 2. Means and standard deviations of Whites and Nonwhites on the Validity Only and Low AI tests.

	Validity Only		Low AI	
	Mean	SD	Mean	SD
Electricians				
White	28.26	10.74	25.17	8.43
Minority	19.79	10.88	23.34	8.80
p	< .01		> .05	
d	.80		.21	
Mechanics				
White	28.56	11.03	18.41	7.21
Minority	20.76	8.74	17.62	6.59
p	< .01		> .05	
d	.74		.11	