Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 3, Number 7, November, 1992 ISSN=1531-7714

Person-Fit Statistics: High Potential and Many Unanswered Questions.

Gerald Bracey,
Consultant

Lawrence M.Rudner,
ERIC Clearinghouse on Assessment and Evaluation

"Whenever we measure anything, whether in the physical, the biological, or the social sciences, that measurement contains a certain amount of chance error....Two sets of measurements of the same features of the same individuals will never exactly duplicate each other....However, at the same time, repeated measurements of a series of objects or individuals will ordinarily show some consistency."

So wrote Robert L. Thorndike in Lindquist's "Educational Measurement" (1951). Traditionally, research into measurement error has dealt with whether or not the test items fit. But over the last 15 or so years, we have seen mounting interest in whether or not the people who answer the items fit. Most of the interest has centered on people whose responses do not fit the typical pattern.

Attempts to systematically identify such people have led researchers to a number of person-fit statistics, with names like caution index, norm conformity index, individual consistency index, and optimal appropriate measurement. This article describes the need for person-fit statistics, summarizes the research on their use, and identifies areas in need of further research.

THE NEED FOR PERSON-FIT STATISTICS

In presenting the need for such statistics, Wright (1977), described several types of people whose response patterns look askew: sleepers who get bored with a test and do poorly on later items, fumblers who do poorly in the beginning because the test format has confused them, plodders who never get to later items, guessers who take wild stabs at the answer, and cheaters.

One can add to the list people who misalign their answer sheets or show exceptional creativity in interpreting questions, people with high ability and atypical schooling, people who do not speak English well, and people who are conservative in their use of partial information.

While personality traits or response styles cause aberrant patterns, Harnisch and Linn (1981) observed that the same number right on a test can mean very different things. On a 20-item test, for example, a score of 10 right can be obtained in 194,756 ways.

Harnisch and Linn also contend that finding aberrant response patterns is no mere academic concern of the psychometrician. They argue that identifying groups or individuals with such patterns can reveal groups with unusual instructional histories or individuals whose test scores cannot be interpreted in standard ways.

When Harnisch and Linn analyzed data from a state testing program, they found that schools in different parts of the state had very different caution indices. They suggest that this result could have been caused by curricula that didn't match the test as well as curricula in other parts of the state. However, the authors give no empirical evidence favoring their conclusion over several others that could also explain the high caution indices.

Tatsuoka and Tatsuoka (1982) offer empirical evidence that patterns of aberrant responses relate to differences in instruction. Two methods were used to teach students addition in signed-number operations. The students then took a test containing both addition and subtraction problems. While the mean differences between the two groups were not significant, the person-fit indices were. In another experiment, students given lessons using different conceptual frameworks showed more aberrance than a group given consistent lesson frameworks.

Frary (1982) has argued that unusual test-response patterns can identify at least some types of test bias. Analysis of scores for properly fitting and poorly fitting examinees, for example, removes some of the noise associated with gross

categorizations such as race or gender.

For the most part, however, person-fit statistics have not yet been applied to many settings. Although the need has been documented and uses suggested, this area has been largely one of potential, not actual, use.

APPLICATIONS OF PERSON-FIT STATISTICS

Rudner (1983) conducted one of the first systematic investigations of person-fit statistics. Using computer-generated data modeling, a well-regarded test, and different types of misfit, Rudner concluded that the statistics have great potential: They can identify significant percentages of examinees with abnormal response patterns. Further, accuracy increases significantly as response patterns become more abnormal.

Papers such as Rudner's have addressed theoretical and methodological concerns about the nature, accuracy, and interchangeability of person-fit statistics. Other researchers have addressed the frequency and amount of abnormal response patterns and, therefore, the practical utility of person-fit statistics.

Frary (1980) calculated four different fit statistics on a large sample of eighth graders who had taken a commercial achievement test battery. Along with the moderate to strong intercorrelations mentioned earlier, Frary found that blacks and females differed from whites and males on some tests. Overall, females showed fewer aberrant responses than males, but racial differences occurred in both directions. Among low-scoring students, the effects were consistent: White and female students made fewer unusual choices.

These findings raised the possibility of test bias. But using a knowledge assurance statistic, Frary concluded that blacks and males did better than whites and females when it came to correctly guessing items for which they had only partial knowledge.

Doss (1981) applied a residual mean-square statistic from the computer program RASCH in the PRIME system to a fifth-grade Chapter 1 setting, where children were given the Iowa Tests of Basic Skills. He examined how removing the poorest fitting 10%, 20%, and 30% of students affected the accuracy of pretest predictions. While the N dropped substantially as he removed students, prediction accuracy increased with each removal.

Although the improvement in accuracy is interesting, Doss's setting may not have provided a meaningful testing situation. The test was badly matched to the students abilities: Even though some (Doss doesn't say how many), took the fourth-grade level of the battery, 25% scored at or below the chance level. After the worst fitting 30% had been removed, only 13% of the Chapter 1 students remained. Finally, the students showed losses from pretest to posttest. The study does, nonetheless, point out another use of person fit statistics--to objectively document whether the testing situation is meaningful.

Schmitt and Crocker (1984) investigated the relationship between scores on the Test Anxiety Scale for Adolescents and person-fit. They used various indices and the Metropolitan Achievement Tests in reading, mathematics, and science in seventh and eighth grades. Students in the middle ability range showed no relationship between test anxiety and person-fit indices. High-ability, low-anxiety students showed greater misfit than high-ability, high-anxiety students. At the low-ability end, the reverse was true: Low-ability, low-anxiety students showed less misfit. The authors offer some conjectures on the findings in terms of a Cognitive-Attentional Theory of Test Anxiety, but present no data that might support their notions.

LIMITATIONS

Two main questions remain for applying fit statistics: (1) Are the statistics theoretically sound? (2) Will they help in practical situations? Some people argue that these basic questions have been answered; others contend that it's too soon to tell.

Person-fit statistics are a logical extension of popular measurement models and thus are well grounded in statistical theory. They are atheoretical, however, with theories of learning and cognition. For example, little research has been done to explain why response patterns may be aberrant. While Tatsuoka and Tatsuoka have looked at consistent errors made by students who apply the wrong mathematics algorithms, causes for the aberrant response patterns listed by Wright and others have not been systematically investigated.

Real questions exist about whether person-fit statistics will make any difference in practical situations. When test takers are unlike the norming group, will their response patterns be sufficiently atypical? How much of a deviation in response patterns is normal for different subject areas in different grades? Are fit statistics sensitive enough to detect strange response patterns when they exist?

The research to date has shown that people with very strange response patterns are indeed detected with few, if any, false identifications. Proponents argue that this is enough to justify routine use of this statistical tool.

REFERENCES

Doss, D.A. (1981) Will removing a few bad apples save the barrel? Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA, April 13-17.

Drasgow, F., M. V. Levine, & M. E. McLaughlin. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, 11(1), 59-79.

Frary, R.B. (1982) A comparison of person-fit measures. Paper presented at the annual meeting of the American Educational Research Association, New York, March 19-23.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18(3), 133-146.

Levine, M., & Drasgow, F. (1982). Appropriateness measurement: Review, critique, and validating studies. *British Journal of Mathematical Statistical Psychology*, 35, 42-56.

Rudner, L. M. (1983). Individual assessment accuracy. Journal of Educational Measurement, Fall.

Schmitt, A. P. and L. Crocker (1984), The relationship between test anxiety and person fit measures. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 23-27.

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics*, 7(3), 215-231.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97-115.

Descriptors: Cognitive Processes; *Error of Measurement; *Goodness of Fit; *Individual Differences; *Learning Theories; Personality Traits; *Research Methodology; Research Needs; Response Style (Tests); Teaching Methods; Test Bias; Theory Practice Relationship

Citation: Bracey, Gerald & Rudner, Lawrence M. (1992). Person-fit statistics: high potential and many unanswered questions. *Practical Assessment, Research & Evaluation*, 3(7). Available online: http://PAREonline.net/getvn.asp?v=3&n=7.