DOCUMENT RESUME

ED 273 665 TM 860 514

AUTHOR	Hambleton, Ronald K.; Swaminathan, H.
TITLE	A Look at Psychometrics in the Netherlands.
PUB DATE	Apr 85
NOTE	12p.; Paper presented at the Joint Annual Meetings of the American Educational Research Association and the National Council on Measurement in Education (Chicago, IL, March 31-April 4, 1985).
PUB TYPE	Information Analyses (070) -- Viewpoints (120) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Bayesian Statistics; Decision Making; Elementary Secondary Education; Foreign Countries; *Generalizability Theory; *Item Analysis; *Latent Trait Theory; Literature Reviews; *Psychometrics; Test Bias; Test Items; *Test Theory
IDENTIFIERS	Netherlands; Rasch Model

ABSTRACT

Comments are made on the review papers presented by six Dutch psychometricians: Ivo Molenaar, Wim van der Linden, Ed Roskam, Arnold Van den Wollenberg, Gideon Mellenbergh, and Dato de Gruijter. Molenaar has embraced a pragmatic viewpoint on Bayesian methods, using both empirical and pure approaches to solve educational research problems. Molenaar presented a taxonomy of Bayesian procedures. Current Dutch research involves nonparametric Bayesian procedures, formalization of prior belief, reporting of the results, and evaluation applications. Van der Linden listed a wide array of testing problems to which decision theory is being applied in the Netherlands. De Gruijter and Mooijaart have made important contributions in least squares Bayesian estimation, and Lewis has clarified difficulties in implementing the hierarchical Bayesian model. Researchers testing Rasch Model assumptions include Van den Wollenberg and Roskam and Jansen. Recent item bias research by Mellenbergh provides a sound method for making inferences about differential item performance between groups. De Gruijter discussed a number of useful applications of generalizability theory, including criterion referenced tests, cutting scores, analysis of ratings, equated scores, and test construction. (GDC)

A Look at Psychometrics in the Netherlands

Ronald K. Hambleton and H. Swaminathan
University of Massachusetts, Amherst

Abstract

In this paper the authors comment generally on advances by Dutch psychometricians in five areas: Bayesian methods, applications of decision theory to testing problems, theory and applications of item response models, item bias research, and uses of generalizability theory.

A Look at Psychometrics in the Netherlands

Ronald K. Hambleton and H. Swaminathan
University of Massachusetts, Amherst

We were pleased to be able to invite a prominent group of Dutch psychometricians to present a symposium at the 1985 Joint Annual Meetings of the American Educational Research Association and the National Council on Measurement in Education in Chicago. Participants were Ivo Molenaar, Wim van der Linden, Ed Roskam, Arnold Van den Wollenberg, Gideon Mellenbergh, and Dato de Gruijter. The focus of these invited papers was recent developments in test theory in the Netherlands. The reason for our invitation was simple: Dutch psychometrics is having a substantial world-wide impact on the development and use of educational and psychological tests. American researchers would benefit considerably from the opportunity to hear more about Dutch psychometrics and to meet some of the world's best-known Dutch psychometricians. Their participation at the AERA/NCME Meetings would contribute positively to the growth and uses of psychometric models and procedures around the world.

In this paper our purpose is to comment generally on the five review papers, and to discuss the significance of the Dutch work for the field of psychometrics.

Bayesian Methods in Dutch Educational Research

The position taken by the Dutch psychometricians with respect to Bayesian methods appears to be eminently sensible. Rather than be embroiled in the philosophical controversies that have racked the statistical world, Molenaar and the Dutch researchers have embraced an eclectic viewpoint that is tempered by pragmatism. This position has resulted in judicious applications of Bayesian procedures to problems that stand to gain by such an approach. Thus, empirical Bayes procedures as well as "pure" Bayesian procedures have been used to solve educational research problems in the Netherlands.

The quantity of Bayesian research in the decade that has elapsed since Melvin Novick "introduced" Bayesian methods to Dutch psychometricians, is staggering. Equally impressive is the breadth of the applications of Bayesian procedures. Bayesian methods have been applied in item response theory, criterion-referenced measurement, linear models, individualized instruction, factor analysis, and evaluation/research methods. It is interesting to note that Molenaar, in presenting his paper, has made a basic contribution to the taxonomy of Bayesian procedures by classifying the procedures according to the nature of prior specifications. Given the role of prior information, this classification scheme is indeed natural and clever.

While some of the Bayesian applications are well-known in the U.S., considerable original research that is not well-known is also being carried out in the Netherlands. This includes non-parametric Bayesian procedures, formalization of prior belief, reporting of the

information contained in a Bayesian analysis, and applications to evaluation. Unfortunately, most of these important works are not available in English, and hence it will take some time for these ideas to be accepted routinely by researchers in the U.S. It should be noted that this point applies to many research papers described in the other four review papers as well.

Applications of Decision Theory to Testing Problems

Formulating testing questions within a decision-theoretic framework was one of the most important psychometric advances of the 1970s. This switch has resulted in the development of new theory for estimating ability scores, determining test lengths, making decisions, and assessing reliability and validity. With the switch to a decision-theoretic framework, test users are forced to consider the decisions they desire to make and the consequences of making mistakes in classifying examinees (for example, failing masters or passing non-masters). Among other things, the decision-theoretic approach draws attention to the problem of setting standards or cut-off scores for the purpose of making decisions (see for example, van der Linden & Mellenbergh, 1977, 1978). Solutions for setting standards are among the most controversial in American testing today.

Despite the controversies, decision-theoretic approaches for testing problems are generally the best approaches today and Dutch psychometricians have been among the most influential contributors to this line of research and development. Professor van der Linden

provides an impressive list of testing problems to which decision theory is being applied in the Netherlands. In the United States, only the applications of decision theory to making mastery classifications in objectives-based instructional programs and making personnel selection decisions are receiving much attention from psychometricians. Both the organizational framework of test uses provided by Professor van der Linden and his comprehensive review of Dutch research in relation to each test use are of substantial value to test users, and will serve to facilitate additional decision theory applications and research. Professors van der Linden and Mellenbergh, and their colleagues, have established themselves as leaders in the world in applying decision theory to testing problems.

Theory and Applications of Item Response Models

The important contributions made by the Dutch psychometricians are in the areas of parameter estimation, and in the testing of the Rasch model assumptions. The contributions to parameter estimation parallel those that have been made in the U.S. particularly with respect to Bayesian estimation (Swaminathan & Gifford, 1982). However, the works of de Gruijter and Mooijaart in the area of least squares Bayesian estimation, and that of Lewis in clarifying some of the difficulties encountered in the implementation of the hierarchical Bayesian model are noteworthy. These results are not well known particularly in the U.S. and hence need further elucidation and dissemination.

The procedure developed by Van den Wollenberg for testing the fit of the Rasch model is especially important since this procedure must be considered as providing a significant improvement over those that are currently available. However, a distinction appears to have been made between the assumption of stochastic independence and unidimensionality. When the latent space is complete and unidimensional , the two assumptions are equivalent. The distinction between these assumptions, if any is implied, is not made clear. Despite this minor quibble, the use of $Q_{1(r)}$ and $Q_{2(r)}$ statistics together may provide the correct test of fit for the Rasch model. The test statistic is asymptotic and hence the sensitivity of the statistic to sample size and test length needs to be studied. A further problem that arises is the feasibility of the procedure when the number of items is large since every pair of item needs to be analyzed. Van den Wollenberg's 'splitter-item' technique for testing unidimensionality shows promise but may be difficult to implement if every item is examined.

Last but not least, the clarification of the nature of measurement and that of the Rasch model provided by Roskam and Jansen is noteworthy. The Rasch model is indeed a useful model. While the Rasch model is the only model that satisfies the requirement of specific objectivity, it is limited in its applications. The issue of the importance of the concept of specific objectivity in comparison to that of generality and utility needs to be looked at. The insight gained through the examination of the Rasch model may provide the

Dutch psychometricians the machinery to deal with the two- and the three-parameter models. There is some evidence of a breakthrough in this arena in the Netherlands. We can only wait and hope for more.

Item Bias Research

Research on the identification of test items which may be unfair to particular sub-groups of examinees such as females, Blacks, or Hispanics has received considerable attention from American psychometricians for about the last ten years (for a review, see Berk, 1982). This interest is not surprising when the importance of the use of test results in the U.S. is considered: Test results are being used among others (1) to place children in special education programs, (2) to influence promotion decisions of children from one grade to the next, (3) to award high school graduation diplomas, and (4) to influence college admissions. Not surprising then, in view of the wide use of tests, questions about their fairness have been raised. Typically, "biased" items are identified by studying the performance of some sub-groups of interest (i.e., Blacks) on subsets of items. Seldom is there interest in learning about the reasons for the malfunctioning test items. If performances of the subgroups differ on the items, then the test items are labelled "biased" and are removed from the test. In fact, as a result of a recent court case in the U.S., one large U.S. test publisher has agreed to only use test items that show no difference in performance between Blacks and Whites. It

matters not that a test item may be revealing of some differential training between the two groups. Differences will not be tolerated. Labelling all such malfunctioning items as "biased" seems wrong to us and may result in lowering the usefulness of test results.

In the U.S., too much time has been wasted in attempting to identify the "best" statistical procedures. Almost no work has been done on the problem of understanding the factors which contribute to item bias.

We view some of the recent research by Professor Mellenbergh and his colleagues as representing the proper direction for future item bias research. Mellenbergh's goal is not only to detect potentially flawed test items (van der Flier, Mellenbergh, Ader, & Wijn, 1984) but to try to understand reasons for these apparently malfunctioning test items. His recommendations for the use of experimental and quasi-experimental designs so that inferences about potential causal variables of differential item performance between groups can be made are sound, and will lead to more understanding about the nature of item bias in tests. The Dutch item bias research therefore is clearly on a constructive course. It should influence the general direction of item bias research in other countries as well.

Uses of Generalizability Theory

Unlike the previous four review papers in which Dutch methodological developments in test theory were highlighted, Dr. de Gruijter focused in his paper on the many applications in the

Netherlands of generalizability theory. Perhaps surprisingly, while most of the relevant theory has been developed in the U.S. (see, Cronbach et al., 1972) there have been relatively few applications. de Gruijter describes useful applications to criterion-referenced measurement, setting cut-off scores, analysis of ratings data, score equating, and test development. Perhaps the main point to be gained from de Gruijter's paper is that many more testing problems than previously known can be viewed within a generalizability framework. We will wait to see whether this new framework for describing testing problems leads to promising solutions, but the prospects are good that it will.

Summary

The contributors to this special issue have done a superb job in organizing the contributions of Dutch psychometricians in five major strands of test theory research. Dutch psychometricians has been immensely successful, especially in recent years, in developing psychometric theory, and in applying psychometric theory to solve a wide variety of educational and psychological testing problems. A review of the references in their papers highlights the fact that they are not working independently of researchers in other countries. Still, it is very clear now that there is a large body of Dutch theoretical and applied testing results that are influencing the testing communities in many countries, including our own. We might add that there appears to be an enthusiasm for psychometric knowledge,

a focus on important problems, and a spirit of cooperation among researchers that sets Dutch psychometric research apart form the work going on in many other countries.

Their work has brought this outstanding group of scholars to the forefront of their field and now the rest of the psychometric world is looking to the Netherlands as one of the centers of excellence for psychometric research.

Dutch psychometrics and Dutch psychometricians are in an enviable position. A handful of dedicated researchers have taken on the problem that plague psychometricians. They have demonstrated that by approaching the problems with a comprehensive long range plan of attack and using technical skills and cooperation among the universities and researchers, as tools, significant progress can be achieved. The Dutch government should be congratulated in having the foresight to support research activities of this nature.

References

Berk, R. A. (Ed.). (1982). Handbook of methods for detecting test bias. Baltimore, MD: The Johns Hopkins University Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements. New York: Wiley.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Dordrecht, The Netherlands: Kluwer-Nijhoff Publishing Co.

Mellenberg, G. J. (1984). Finding the biasing trait(s). Paper presented at the Advanced Study Institute: Advances in Measuring Cognition and Motivation, Athens, Greece.

Swaminathan, H., & Gifford, J. A. (1982), Bayesian estimation in the Rasch model. Journal of Educational Statistics, 7, 175-192.

van der Linden, W. J., & Mellenbergh, G. J. (1977). ~~Optional~~ Optimal cutting scores using a linear loss function. Applied Psychological Measurement, 1, 593-599.

van der Linden, W. J., & Mellenbergh, G. J. (1978). Coefficients for tests from a decision theoretic point of view. Applied Psychological Measurement, 2, 119-134.

van der Flier, H., Mellenbergh, G. J., Ader, H. J., & Wijn, M. (1984). An interactive item bias detection method. Journal of Educational Measurement, 21, 131-145.