REVISION DE LIBROS/BOOK REVIEW

EDUCATIONAL MEASUREMENT

Linn, R.L.(Ed.)(1989,3rd ed.).New York: American Council On Education. Macmillan Publishing Company (610 pp).

The third edition of the volume Educational Measurement gives, as the previous two editions of Lindquist (1951) and Thorndike (1971), a comprehensive review of the state of art of educational measurement. The volume is edited and introduced by R.L. Linn and is organized in three parts:(1) Theory and General Principles (chapters 2 through 7), (2) Construction, Administration, and Scoring (chapters 8 through 11), and (3) Applications (chapters 12 through 18). More than half of the number of pages is devoted to theory and general principles and the emphasis of the review is also on this part.

Chapter 2, Validity by S. Messick: Traditionally, content, criterion-related, and construct validity are distinguished. Messick, however, emphasizes that validity is a unified concept and that also social consequences of testing must be included in the concept. He distinguishes two facets of the validity concept. First, source of justification of testing: evidential versus consequential basis of testing. Second, the function of testing: test interpretation versus test use. The crossing of the two facets yields a four-fold classification system of the validity concept. First, the combination of evidential basis of testing and test interpretation is construct Validity,i.e. the evidence that supports the interpretation of test scores in terms of concepts and theoretical and empirical relations with other concepts. Second, the combination of the consequential basis of testing and test interpretation yields the value implications of tests, such as the consequences of ideological values on test interpretations. Third, the combination of evidential basis of testing and test use is also construct validity with specific emphasis on applications and utility. Finally, the combination of consequential basis of testing and test use is also construct validity with specific emphasis on applications and utility. Finally, the combination of consequential basis of testing and test use is the concern with the social consequences of testing. For each of the four elements of the classification an extensive overview of problems, concepts, and methods is given. Messick's claim that validity is a unified concept and that social consequences of testing must be considered are essentially correct. The distinction between the traditional concepts of content, criterion-related, and construct validity is usually hard to make; the inclusion of the social consequences is a rather logical step in the development of the validity concept. But, there is also a real danger in the broadening of the validity concept: The concept might become too broad and complicated for practical use in applications. In my opinion it is time to consider whether a rat-

Psicothema, 1992, vol. 4 nº 1, pp. 317-321
ISSN 0214 - 9915 CODEN PSOTEG

her simple definition of essential aspects of validity and guide lines for validation research can be developed for practical applications.

Chapter 3. Reliability by L.S. Feldt and R.L. Brennan. The chapter discusses reliability in classical test theory, generalizability theory, and criterion-referenced testing. It gives a comprehensive review of reliability in classical test theory and generalizability theory; concepts, coefficents, estimation procedures, and applications in special situations are thoroughly discussed. The part on reliability for criterion-referenced tests, however, contains a conceptual mistake. In classical test theory reliability is defined as the squared correlation of observed and true score; this squared correlation equals the correlation of two parallel measurements. Analogously, Feldt and Brennan consider the reliability of a criterion-referenced test as the association of pass/fail decisions on two parallel measurements. But, Mellenbergh and Van del Linden (1979) showed that the association of pass/fail decisions of two parallel measurements is not related to the association of the true master/nonmaster dichotomy and the observed pass/fail dichotomy in the same way as in classical test theory. Feldt and Brennan discuss pass/fail decision consistency, but the consistency of decisions is not related in the same way to the association between pass/fail decisions and master/nonmaster true states as consistency of measurements is related to the reliability coefficient.

Chapter 4. Principles and Selected Applications of Item Response Theory by R.K. Hambleton. Item response theory is one of the major developments in educational and psychological testing. The chapter is a well-written and comprehensive introduction in item response theory. It treats item response theory in the same way as Hambleton and Swaminathan's (1985) book. The chapter discusses the assumptions and main models of item response theory, item and test information functions and efficiency, and parameter estimation procedures. A nice feature of this chapter is the emphasis on methods for investigating the fit of models to empirical data. In applications of item response theory often large samples are used, which means that the use of statistical tests is of limited value; other types of cheques on model fit are needed and they are extensively treated in this chapter.

Next to conceptual issues Hambleton also discusses important apllications, such as test development, item bias, and adaptive testing. I have a rather personal comment, which applies to all treatments of item response theory such as Hambleton and Swaminathan'ns (1985) book and Muñiz's (1990) Spanish introduction. The authors make the impression that item response theory is a rather isolated and independent development that is not related to other types of models. Actually, item response theory is related to other latent trait models such as factor analytic models; see, for example, Goldstein and Wood (1989).

Chapter 5. Bias in Test use by N.S. Cole and P.A. Moss. The concept of bias of test use is discussed within the framework of the validity concept. Bias is considered to be differential validity of test score interpretation between relevant subgroups of a population. Five Categories are distinguished: (1) constructs in context, (2) content and format, (3) test administration, (4) internal test structure, and (5) external test relations. In each of these categories evidence about bias can be examined. In the category Content and Format, for example, judgmental methods to examine item content on stereotyping are classified. In the Internal Test Score category item response theory based methods and their approximations to investigate item bias are discussed. Under the heading External Test Relations differences in predictor criterion relations between groups are described. The chapter ends with a discussion of extra-validity aspects such as value and policy issues in testing. The chapter gives a good review of the bias concept and the methods for investigating item and test bias. Since the seminal paper of Patersen and Novick (1976) it is clear that may value-issues of testing can be discussed within the context of psychometric decision theory. In the chapter decision theory is briefly addressed, but in my opinion this topic deserves more attention.

Chapter 6. Scaling, Norming, and Equating by N.S. Petersen, M.J. Kolen, and H.D. Hoover. Scaling is the process of assigning numbers to examinees'test performance. Scales can convey specific information, e.g. the transformation of raw scores to a normal distribution contains information on examinees'relative position in the group of examines'relative position in the group of examinees, while scale intervals contain information on measurement precision. Norms are the basis for comparing an examinees'performance to the performance of a relevant group of examinees. Equating is the process of assigning numbers to different test forms such that they are equivalent in some sense. The condition for equated scores can theoretically be defined but, unfortunately, all conditions cannot simultaneously be fulfilled. Therefore, equating methods are approximations. The authors give a useful overview of data collection methods and of statistical methods for equating scores.

Chapter 7. Implications of Cognitive Psychology for Educational Measurement by R.E. Snow and D.F. Lohman. In this chapter developments in cognitive psychology are discussed, mainly to introduce measurement specialists into cognitive psychology. From the point of view of cognitive psychology measurement models have three flaws. First, item performance is not explained in psychological terms. Second, measurement models make unrealistic and simplistic assumptions. Third, test validation is concerned with external evidence and is no part of the measurement model. Recently, attempts are made to integrate cognitive psychological and measurement models. Examples are the explanation of item difficulty parameter in item response models by incorporating cognitive information processing models and the construction of tests according to substantive facet desings. The authors advocate the further integration of measurement and cognitive models.

Chapter 8. The Specification and Development of Tests of Achievement and Ability by. Millman and J. Greene The chapter gives an overview of item and test construction methods: defining test purpose, specifyng the test, developing items, evaluating and analyzing items, selecting the items for the test, and assembling the tests. The content of this chapter is traditional and is well-known in the literature. The chapter gives a good overview of the first three topics (test purpose, test specification, and item development); the content is traditional but there seem to be not many new develpments in this area. The treatment of the last three topics is, however, old fashioned. Item analysis is completely considered from the classical point of view and for modern methods the reader is referred to Hambletons'chapter on item response theory. The classical methods are treated in introductory textbooks, such as Croker and Algina (1986), and the authors miss the opportunity to show how item response theory methods for item analysis can be incorporated in the test construction process. Moreover, in a note at the end of the chapter the authors state that the chapter was writtren in 1985. Therefore, they miss the most important recent development in item selection and test assembly. In 1985 Theunissen published the first paper on linear programming techniques for designing tests and since that year many papers obn that topic appeared in the psychometric literature. The chapter should have been revised before it was published in 1989.

Chapter 9. The Four Gnerations of Computerized Educational Measurement by C.V. Bunderson. D.K. Inouye, and J.B. Olsen. The authors distinguish four generations of computerized measurement. The first one is computerized testing, where conventional tests are administered by a computer. The second generation is computerized adaptive testing. The item difficulty and content are adapted to the subject; item response theory is used as a tool for adapting the items to the subjet. In the third generation measurement is integrated with the curriculum and students'performance is unobtrusively measured and monitored during the learning process, such as in mastery assessment systems. The fourth generation is intelligent measurement, where intelligent scoring, interpretation of scores and advices to students and teachers are applied.

Chapter 10. Computer Technology in Test Construction and Processing by F.B. Baker. The introduction of computer technology is the major development in educational measurement. Baker discusses the use of microcomputers in the whole process of educational measurement: item writing, item banking, test construction, test printing, and reporting of results. An example of an integrated computer system is MICROCAT. The author also discusses the important role of item response theory in this type of systems.

Chapter 11. The Effects of Special Preparation on Measures of Scholastic Ability by L. Bond. The chapter gives an overview of empirical studies of the effects of special preparation, or coaching, on scholastic aptitude performance. Coaching has a positive effect but it is rather small: a gain of about one or two correctly answered items on the Scholastic Aptitude Test. Moreover, the gain is nonlinearly related to the contact time of special preparation: the gain is becoming smaller for increasing contact time.

Chapter 12. Designing tests that are integrated with instruction by A.J. Nitko. Instructional and test design are related and integrated processes. Four categories of instructional design are distinguished: placement, diagnotic, monitoring, land attainment decisions. The categories need different types of tests and they can be used to design tests.

Chapter 13. Administrative Use of School Testing Programs by J.A. Frechtling. School administrators have used test scores in three different ways. Traditionally, test scores are used for reporting students'achievement. Test scores are also used for evaluating the effectiveness of educational programs. Finally, test scores are increasingly used as an indicator how well an educational institution is functioning.

Chapter 14. Certification of Student Competence by R.M. Jaeger. Educational institutions use tests to decide whether students will be certified of their minimum competence, e.g. by giving them a high school diploma. In competency testing test scores must be evaluated against a predetermined standard for minimum competence. A review of standard setting methods is given. From the literature it is clear that different standard-setting methods yield a wide variety in standards. The median ratio of the largest and smallest standard determined with two different standard setting methods (computed over 32 comparison of methods) appeared to be 1.46, e.g. one method sets a standard for minimum competence at 20 correctly answered items, whereas another method sets it at 1.46 x 20 = 29 correctly answered items for the same test. Another question is whether the students had the opportunity to learn and were provided with adequate instruction in the skills that are tested. For that reason competency testing was also challenged in court in the United States of America.

Chapter 15. Educational admissions and Placement by D.R. Whitney. In education tests are used for two main types of decisions: (1) admission of students to educational institutions, and (2) placement of students in courses best suiting their skills. In this chapter the use of admission and placement tests is discussed.

Chapter 16. Counseling by L.W. Harmon In counseling tests are used to investigate whether the individual has (1) appropriate background knowledge and skills to begin an educational program, and (2) the ability to complete the program.

The issues of test validity and culturally biased assumptions as well as computerized methods are discussed.

Chapter 17. Identification of Mild Handicaps by L.A. Shepard. The chapter discusses the identification of mildly handicapped children. Psychological test are not very successful in identifying these children, not because of low validity of the tests but because of the low base rate of mildly handicapped children in the population. For the purpose of research the emphasis must be on the identification os specific subpopulations. Researchers can exclude from their studies children that do not meet strict criteria, which is usually not done in the practice of assignment of children to special schools or special educational programs. A policy conclusion is that quota restrictions should

be placed on special education because of the tendency to assign more and more children to special educational programs.

Chapter 18. Testing of Linguistic Minorities by R.P. Duran The chapter gives an overview of empirical research in fire areas: (1) language proficiency assessment, (2) cognitive assessment, (3) assessment of school achievement, (4) special education assessment, and (5) assessment for college admissions.

As the previous editions of Linguist (1951) and Thorndike (1971) the third edition of Educational Measurement contains a wealth of information on recent developments. Some points are stressed. First, the most spectacular recent development is the integration of computer technology in educational measurement. The psychometric concepts of item response theory were known for 40 years (Lord, 1952) but only computer technology made it possible to apply the theory at large. Computer technology combined with item response theory leads to complete new ways of constructing, administering, scoring, analyzing, and applying tests. Second, the less techinical areas, such as item writing and the application of cognitive psychological models, develop rather slowly. Third, at least in the United States of America the number of legal issues in testing is rapidly increasing.

The format of this volume does not tempt to reading: It measures 28 x 22 cm, has more than 600 pages, and is printed in two columns. Nevertheless, educational researchers, students, and policy makers are highly recommended to study the book.

REFERENCES

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42,* 139-167.

Hambleton, R.K., & Swaminthan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Lindquist, E.F. (Ed., 1951) *Educational Measurement* Washington, D.C.: American Council on Education.

Lord, F.M. (1952). A Theory of test scores. *Psychometric Monograph*, No. 7.

Mellenbergh, G.J., & van der Linden, W.J. (1979). The internal and external optimality of decisions based on tests. *Applied psychological Measurement, 3,* 257-273.

Muñíz, J. (1990). *Teoría de respuesta a los ítems*. Madrid: Ediciones Pirámide, S.A.

Peterson, N.S., & Novick, M.R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13,* 3-29.

Theunissen, T.J.J.M. (1985). Binary programmning and test design. *Psychometrika, 50,* 411-420.

Thorndike, R.L. (Ed., 1971). *Educational measurement 2nd ed).* Washington, D.C.: American Council on Education.

Reviewed by:

Gideon J. Mellenbergh: Faculty of Psychology, University of Amsterdam, the Netherlands/Interuniversity Graduate School of Psychometrics and Sociometrics, the Netherlands