The Changing Conception of Measurement: A Commentary

Ronald K. Hambleton
University of Massachusetts

This paper comments on the contributions to this special issue on item banking. An historical framework for viewing the papers is provided by brief reviews of the literature in the areas of item response theory, item banking, and computerized testing. In general, the eight papers are viewed as contributing valuable technical knowledge for implementing testing programs with the aid of item banks.

As noted by van der Linden (1986), measurement models and testing practices have undergone some major changes in the last 20 years. One important change is the shift from the nearly exclusive use of classical test models and standardized tests to the use of newer test models derived from item response theory (IRT) and customized tests (i.e., tests that have been constructed to match user needs). Test administration procedures have also changed: Group-administered paper-and-pencil tests are sometimes replaced by tests administered adaptively at computer terminals. There has also been a major shift in emphasis from norm-referenced to criterion-referenced test score interpretation and a concomitant desire in the testing field to tie testing and instruction more closely together.

There are signs of change throughout the testing industry. Large-scale testing programs such as Educational Testing Service's Scholastic Aptitude Tests use IRT models to equate test forms. Major test publishers such as CTB/McGraw-Hill and The Psychological Corporation are using IRT models to develop tests and to equate test scores. They are also using IRT models to help produce customized standardized achievement tests. Departments of education in states such as California and Maryland are using IRT models in their state assessment programs in reporting scores and addressing item bias. The College Board's popular Degrees of Reading Power, which links test results to reading instructional levels, uses IRT as the measurement model to accomplish its goals. School districts such as Portland, Oregon and Los Angeles, California are using item banks and achievement scales which are based on IRT models to deliver tests and score information to schools and homes. The National Assessment of Educational Progress, a major federal testing program aimed at helping policymakers monitor the quality of education in the U.S., is using IRT models in scaling scores and reporting information. Computerized adaptive testing is now in use or under serious consideration by the U.S. Armed Services and several of the large medical and allied health organizations.

Other important signs of change in testing in the U.S. include the major shift in emphasis of state and school district testing from *norm-referenced* to *criterion-referenced*. Nearly every state in the U.S. now supports at least one major criterion-referenced testing program. Over 800 organizations (e.g.,

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 4, December 1986, pp. 415–421

0146-6216/86/040415-07$1.60

medical, allied health, insurance) have initiated certification, recertification, and/or licensing exams for assessing professional competencies and awarding certificates, licenses, and the like. Also, school districts across the country are experimenting with computers for test storage, construction, scoring, and reporting, and item bank construction has become a popular topic in both large and small school districts (Brzezinski & Hiscox, 1984).

In summary, the education and measurement journals are filled with articles addressing important changes in measurement practices. These changes are also manifesting themselves in actual testing programs. Among the most promising changes are those that concern item banking and related topics. This special issue of *Applied Psychological Measurement* brings together a distinguished international group of scholars who have been contributing to the advancement of new measurement theory, item banking procedures, and computerized testing.

The history of item response theory can be traced to the work of Richardson, Lawley, Tucker, and others in the 1930s and 1940s. Although van der Linden (1986), in his lead article, associates Binet's work on IQ testing with classical test theory and standardized testing, Binet himself introduced the notion of the item characteristic curve, which is the most important concept in IRT. Important research and publications by Lord (1952), Rasch (1960), and Wright (1968) led the way for an avalanche of technical reports, papers, and workshops that have provided the technical base of knowledge for many of the present-day IRT applications. Some of the promising applications described in this issue include item banking, test development, test score equating, and adaptive testing.

The property of greatest interest in these applications concerns "invariance." When the IRT model of choice fits the test data, item and person parameter estimates are invariant in the sense that neither the choice of examinee sample from the examinee population of interest nor the choice of test items from the item population will affect the item and ability parameters (though the standard errors of the parameter estimates may be affected). But problems continue to impede the use of IRT results and applications in actual practice, including parameter estimation difficulties, scarcity of fast and uncomplicated computer programs, and the lack of guidelines for determining model fit (Hambleton & Swaminathan, 1985).

Like IRT, the concept of item banking has also attracted considerable attention from public and private agencies. The current interest in item banking is actually the second serious effort to make the concept work. The first effort in the late 1960s and early 1970s in the U.S. and Great Britain (see, e.g., Wood & Skurnik, 1969) failed because of the tremendous amount of paper, administration, and organization necessary to implement item banking. The availability of computer software for storing and retrieving test items, for facilitating item revisions, and for printing tests has enhanced the value of item banks considerably (Brzezinski & Hiscox, 1984).

The concept of item banking has been closely connected in the past with the behavioral objectives and individualized instruction movements which have flourished in the U.S. since about 1965. When item banks are stocked with valid test items, the task of preparing top-quality tests in the schools, for example, is made easier and the product is usually better than could be produced without the item bank.

The third and final cornerstone on which most current measurement practices are based involves the computer. The current success of IRT and item banking is directly tied to the role of computers (both mainframe and microcomputers). Parameter estimation, equating of test scores, item storage, item retrieval, test printing, and test scoring and reporting are among the most popular and important uses of the computer.

These times are important for measurement specialists because the technical problems associated with item banking, applications of IRT, and computerized testing must be solved in a period of high expectations for what tests can do in education, industry, and the armed services. Practical solu-

tions to current testing problems are in great demand because of the central role that tests play in our society.

The contributions to this special issue successfully address many of the technical problems associated with current testing practices and, therefore, the number of obstacles to proper implementation of item banks has been reduced considerably. The papers fall roughly into three broad categories: (1) establishing the item bank—item development (Hornke & Habon), new item types (Masters & Evans), parameter estimation (van der Linden & Eggen), and linking items to a common scale (Vale); (2) test design from item banks (Theunissen); and (3) delivery system design (van Thiel & Zwarts) and item banking issues (Baker).

Establishing the Item Bank

Preparing sets of valid test items to measure content domains of interest is a central problem in any customized testing project such as item banking. Well-developed item writing and item validation methods are readily available for application to multiple-choice items (see Hambleton, 1985; Popham, 1978; Roid & Haladyna, 1982), although they are not always used. Also, algorithms for generating test items are attracting interest and appear to be a way to produce equivalent multiple-choice test items that are closely matched to the content domains of interest (Millman, 1984). But the item writing field is weak at present in the generation of new item formats to meet the measurement needs of practitioners and to enhance test score validity. This special issue offers two promising hopes for the future: theory-based test items (Hornke & Habon, 1986) and non-dichotomously scored test items (Masters & Evans, 1986).

Theory-based test item construction is a new topic encouraged and directed by cognitive psychologists in the U.S. such as Embretson (1984, 1985), Glaser (1981), and Sternberg (1981), and in Europe by Fischer (1976) and several of his students. The notion is that test items can be constructed to reflect precisely the cognitive processes which are of interest to test developers. Often, these processes come from a careful analysis of a set of tasks which describe some problem-solving domain. Theory-based test item writing represents a merger of cognitive theory and psychometrics and, as such, represents one of the most promising future directions for testing.

Hornke and Habon (1986) not only were able to highlight the construction of theory-based test items to measure abstract reasoning (in the manner of Raven's Progressive Matrices Test) along with the development of a method for identifying underlying cognitive processes in their research, but they were also able to demonstrate the utility of the linear logistic test model (Fischer, 1976) to account for examinee item performance. IRT models like those of Fischer (1976) and Embretson (1984) will become more popular as the work of cognitive psychologists in the area of test development problems becomes better known.

Though Hornke and Habon's premise of a single subject matter expert preparing an item bank is almost never seen in practice, the desirability of theory-based item development is clear even if item banks could be put together by hundreds of item writers with extensive item review and pilot-testing capabilities. Quite simply, theory-based item construction requires less extensive pilot testing, validation procedures are more clear-cut, and score interpretations are enhanced. Although the authors are fairly optimistic about their work and the desirability of theory-based item development is obvious, some work currently being done for the U.S. Air Force in constructing cognitive tests suggests that test construction can be slow, tedious, and very labor-intensive (Hambleton, Swaminathan, Arrasmith, Gower, Rogers, & Zhou, 1986). The practical feasibility of theory-based item development is still uncertain.

The Masters and Evans (1986) contribution to item banking is especially important because they expand item banking to include polychotomously-scored items: essay questions, multistep problems, and performance items (scored with a rating scale). Such additions are invaluable if item banks are to

serve the needs of users. But items using the formats introduced by Masters and Evans can, in most cases, be added to an item bank (though, unfortunately, they seldom are included). The major contribution of Masters and Evans, in addition to highlighting the need for new item formats, is to introduce some IRT models that can handle the new item formats so that these new formatted items can be "linked" to other items in a bank (see, e.g., Wright & Masters, 1982). In this way, all of the advantages of a calibrated bank of items can be achieved. Without linking, some desirable features of item banking are lost. The authors provide clear illustrations for how items can be calibrated using the partial-credit IRT model, how other items can be linked, and how the full set of items can be used in adaptive testing.

Masters and Evans feel that the partial credit model, and others like it, are the most important part of an item bank. Certainly the models are important for producing such applications as adaptive testing. However, the concept of item banking has appeal even when the test items are not linked to an IRT ability scale. Properly validated test items can be immensely useful to teachers, for example, in preparing classroom criterion-referenced tests even though the items are not linked to a common ability scale. In fact, linking may be unnecessary for some applications and even undesirable for others because the unidimensionality assumption may not be viable. It would be misleading to convey the impression that measurement models are more important than the test items. For some applications, the statement by Masters and Evans is definitely true, but for others it is not.

Estimating item parameters and placing them on a common scale for test items in a bank is a normal activity for item bank developers. Because there are usually more items in a bank than can reasonably be administered to the available examinee pool, a sampling plan is usually designed by dividing the available items into k randomly-equivalent forms and then administering each form to $1/k$ of the examinees. In addition, some common items are added to each form for the purpose of "linking" all of the test items to a common scale.

Van der Linden and Eggen (1986) have offered an attractive alternative to the usual data collection design. In this alternate design, data are collected over time, using an optimal item selection procedure, to update (revise) item parameter estimates. Such a solution seems especially desirable for developers with a modest-sized examinee population, as is the case in many small school districts. Districts can move into item banking using first approximations to the item parameter estimates. Over time, these estimates can be updated and improved using the van der Linden-Eggen algorithms. A Bayesian approach represents an excellent way to address the problem (see, e.g., Swaminathan & Gifford, 1982, 1986). Though their approach is restricted to the Rasch model, school districts often prefer to work with this model because of its simplicity and the difficulty of obtaining large enough sample sizes to calibrate items using the three-parameter logistic model.

Vale (1986) has provided an excellent review of some of the linking designs and transformation methods in popular use, along with some new results on linking designs. Perhaps the most important findings are that an interlaced design may be the best of the designs, and that two common items per pair of forms may be sufficient to produce good results. Practically speaking, however, anchor test designs are somewhat easier to implement, and the associated analyses are definitely easier to carry out. Therefore, in practice, interlaced designs may be less desirable than they appear. The finding that an anchor of two items is sufficient is encouraging because this is the second time that such a result has been found (see Wingersky & Lord, 1984). Practitioners should be aware that the use of so few common items could lead to grossly misleading results if one item is flawed.

Test Design from Item Banks

Theunissen (1986) has tackled the important problem of test design from an item bank with some very complicated optimization algorithms. The author provides an excellent statistical solution to the test design problem, that is, test construction to fit

a desired target information function. Presumably, when assessments are to be made at the objective level, the algorithms can be repeated for each objective, with an appropriately chosen target information curve for each objective. Guidelines will be needed, obviously, to help test developers determine the curves (or, alternatively, to specify the information at points of interest on the ability scale). An equally common situation arises when test developers have a detailed set of content specifications to guide the test development process. Certainly the specifications can be formulated as a knapsack problem (KP); however, the complications in doing so are best avoided. Providing the percent of test items desired in each cell of the test specifications seems to be about the heaviest demand that can be made of test developers.

Delivery Systems for Item Banks

Van Thiel and Zwarts (1986) have described a support system for the use and maintenance of item banks in schools which is under development in the Netherlands. This state-of-the-art system is an excellent platform for advances in psychometrics and item writing technology. The planned system is considerably more sophisticated than most of the systems now in use in the United States (notable exceptions include the work by WICAT Systems and Assessment Systems Corporation).

Hambleton (1984) described six desirable characteristics of a computer-based test development system:

1. An item bank with easily retrievable items, along with related information such as reference material and item usage data.
2. An objective item pool, with a meaningful classification scheme for the objectives.
3. The capability to search and retrieve items for test development.
4. Automatic generation of tests.
5. Interactive delivery of tests.
6. Analysis of item performance data and automatic storage of the data with the associated test items.

All of these desirable characteristics are present in the van Thiel-Zwarts system, with minor exceptions: Provisions for handling artwork and other graphic materials are not specified; also, current plans do not include an objectives bank. Satisfactory provisions for handling graphic material are essential; otherwise, users may become frustrated and/or stop assessing competencies with needed graphic material and thereby reduce their test validity. The addition of an objectives bank would be easy to accomplish, if desired.

While the design of the van Thiel-Zwarts system is praiseworthy, the next round of problems in their research and development may be more difficult to resolve. Various questions will be central:

1. Which IRT model should be chosen? What is the evidence for model-data fit? What are the practical consequences of model-data misfit for various uses of the item bank? (See, e.g., Divgi, 1986.)
2. Can the system handle new item formats such as those highlighted by Masters and Evans (1986) and Gulliksen (1986)?
3. How easy will the system be to use? As Baker (1986) has noted, teachers, for example, may not use the system unless it is very simple and quick to use.
4. What effect might instruction have on (a) the invariance property of item statistics and (b) the nature of the trait or traits measured by the item bank? (This question is currently worrying many IRT advocates in the U.S.)

Baker (1986) does, in fact, offer a number of reservations about item banking that should be given serious attention by researchers; he correctly notes that test development using item banks may not be as simple as item banking advocates have reported. Baker also points out that teacher interest in item banking and test development may be quite limited at the present time. Possibly he is correct, but there are many examples of successful item banking and test development in several locations in the U.S. that can serve as models for others. With the development of more microcomputer testing software, for example, and more knowledge among test developers about this software, the number of applications should expand tremendously. The cur-

rent situation with item banks is similar to the status of word processors and video cassette recorders several years ago. Little was then known about these devices, but once their capabilities became better known, they became indispensable. Baker's concerns and reservations about item banks are appropriate, but item banks and computers should ultimately solve a variety of testing problems.

Baker introduces the concept of a test bank in his paper. This concept seems useful for some applications, such as the construction of pretests and posttests for well-defined and highly structured individually prescribed instructional programs. On the other hand, there are a great many more applications where too much flexibility in test development is lost when the concept of test banking is adopted. Test banking appears to be a cost-effective special case of item banking that will be useful for meeting only a few special testing needs.

Conclusion

The immense potential of item response theory, item banking, and computers for improving testing practices is generally well-known. What remains to be developed are the precise details for using these new methodologies to solve testing problems. The authors of the papers in this special issue have not solved all of the technical and implementation problems, but they have solved many of the problems and provided directions for solving others. It is hoped that this special issue will focus the attention of other researchers on the remaining sources of difficulty.

References

Baker, F. B. (1986). Item banking in computer-based instructional systems. *Applied Psychological Measurement, 10,* 405–414.

Brzezinski, E. J., & Hiscox, M. D. (Eds.) (1984). Microcomputers and testing [Special issue]. *Educational Measurement: Issues and Practices, 3,* 4–34.

Divgi, D. R. (1986). Does the Rasch model really work for multiple-choice items? Not if you look closely. *Journal of Educational Measurement, 23,* 283–298.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175–186.

Embretson, S. E. (Ed.) (1985). *Test design: Developments in psychology and psychometrics*. New York: Academic Press.

Fischer, G. H. (1976). Some probabilistic models for measuring change. In D. N. M. de Gruijter & L. J. Th. van der Kamp (Eds.), *Advances in psychological and educational measurement*. New York: Wiley.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist, 36,* 923–936.

Gulliksen, H. (1986). Perspective on educational measurement. *Applied Psychological Measurement, 10,* 109–132.

Hambleton, R. K. (1984). Using microcomputers to develop tests. *Educational Measurement: Issues and Practices, 3,* 10–14.

Hambleton, R. K. (1985). Criterion-referenced assessment of individual differences. In C. Reynolds & V. L. Willson (Eds.), *Methodologies and statistical advances in the study of individual differences*. New York: Plenum.

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

Hambleton, R. K., Swaminathan, H., Arrasmith, D., Gower, C., Rogers, H. J., & Zhou, A. (1986). *Development of an integrated system to assess and enhance basic job skills* (Air Force Research Report No. 2). Amherst MA: School of Education, University of Massachusetts.

Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10,* 369–380.

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

Masters, G. N., & Evans, J. (1986). Banking non-dichotomously scored items. *Applied Psychological Measurement, 10,* 355–367.

Millman, J. (1984). Individualizing test construction and administration by computer. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction*. Baltimore MD: Johns Hopkins University Press.

Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs NJ: Prentice-Hall.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Roid, G. H., & Haladyna, T. J. (1982). *A technology for test-item writing*. New York: Academic Press.

Sternberg, R. J. (1981). Testing and cognitive psychology. *American Psychologist, 36,* 1181–1189.

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

Swaminathan, H., & Gifford, J. A. (1986). Bayesian

estimation in the three-parameter model. *Psychometrika, 51,* 589–601.

Theunissen, T. J. J. M. (1986). Some applications of optimization algorithms in test design and adaptive testing. *Applied Psychological Measurement, 10,* 381–389.

Vale, D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10,* 333–344.

van der Linden, W. J. (1986). The changing conception of measurement in education and psychology. *Applied Psychological Measurement, 10,* 325–332.

van der Linden, W. J., & Eggen, T. J. H. M. (1986). An empirical Bayesian approach to item banking. *Applied Psychological Measurement, 10,* 345–354.

van Thiel, C. C., & Zwarts, M. A. (1986). Development of a testing service system. *Applied Psychological Measurement, 10,* 391–403.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347–364.

Wood, R., & Skurnik, L. S. (1969). *Item banking*. London: National Foundation for Educational Research.

Wright, B. D. (1968). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton NJ: Educational Testing Service.

Wright, B. D, & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Author's Address

Send requests for reprints or further information to Ronald K. Hambleton, University of Massachusetts, Hills South, Room 152, Amherst MA 01003, U.S.A.