The Changing Conception of Measurement in Education and Psychology

Wim J. van der Linden
University of Twente
Enschede, The Netherlands

Since the era of Binet and Spearman, classical test theory and the ideal of the standard test have gone hand in hand, in part because both are based on the same paradigm of experimental control by manipulation and randomization. Their longevity is a consequence of this mutually beneficial symbiosis. A new type of theory and practice in testing is replacing the standard test by the test item bank, and classical test theory by item response theory. In this paper it is shown how these also reinforce and complete each other.

The first use of the standardized test in education and psychology is usually connected with the year 1905 and the name Alfred Binet. Binet's contribution was a direct consequence of his membership in an advisory committee to the French Minister of Education, appointed in 1904. The committee was given the task of formulating recommendations that would solve the problem of instruction of retarded children in the Paris schools. In view of this, it was proposed to transfer the least gifted children to special schools where they would be taught a simplified curriculum, and the committee was requested to report on a method to differentiate between children with mental retardation and those who, although able to learn, did not perform well. This decision could not be left to their teachers for fear they would apply their own criteria and be less than objective. In its final report, the committee decided on what is now known as the intelligence test. Binet accepted the assignment to develop the instrument and, together with Th. Simon, produced the Binet-Simon intelligence test (Binet & Simon, 1905), which was the first standardized intelligence test.

Binet's merits as a test researcher can be better understood against the background of the work of such contemporaries as Galton, Wundt, Ebbinghaus, Pearson, and Spearman (see DuBois, 1970). These colleagues were mainly active in fields such as anthropometrics (the systematic measurement and comparison of physical properties of the human body) and psychophysics (the study of the psychological sensation of physical stimuli, e.g., light and sound). In their experiments, these researchers instructed examinees to perform sensory and motor tasks, and measured their performance with the aid of simple devices in units of time, distance, pitch, and power. The measurement involved can be characterized as direct measurement supported by established physical theory.

Binet, on the other hand, was less interested in measurement of physical properties or their sensation. He undertook the task of entering the domain of the purely mental functions in order to define and measure intelligence. The physical law-based measuring devices of the anthropometricians and psychophysicists were of little utility when an

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 4, December 1986, pp. 325–332

0146-6216/86/040325-08$1.65

intangible such as intelligence was to be measured. Binet's solution was to construct a large number of specific tasks (today called test items), all eliciting behaviors indicative of intelligent capacities. The prescribed method of item administration was extensively documented in a set of standard instructions. By trying out the items on various age groups, Binet was able to establish typical performance for each age. Using items normed in this way, he was also able to score test performance on a scale for mental age. Stern (see DuBois, 1970) later proposed to divide mental age by chronological age, resulting in the well-known measure of the intelligence quotient (IQ).

For several reasons Binet's contribution was innovative in character. Three new perspectives opened by him are briefly discussed here. First, Binet constructed a test consisting of a large number of separate items. In all, the first version of his test consisted of no fewer than 30 subtests for various mental abilities (e.g., the immediate reproduction of figures, naming objects, ranking various quantities). Binet's approach thus differed decisively from that used in the anthropometric or psychophysical experiments, which usually dealt with examinee responses to a single task, primarily in order to vary certain stimuli in intensity.

There were two good reasons for this *test lengthening*. The first concerned the issue of test validity. Unlike his contemporaries, Binet was attempting to measure a complex human ability, and to do justice to this a large collection of tasks or items was needed. In fact, the ideal set of items in his test would constitute an abstract representation of the large variety of problems which may confront a person in everyday life and for the solution of which intelligent behavior is required. A smaller number of items in the test would mean less representativeness. Binet's second consideration would today be termed "reliability." It is interesting to note a parallel between his conception of the intelligence test and the theoretical work of Charles Spearman who, building on the pioneering work of Francis Galton and in close association with the mathematical work of Karl Pearson, formulated the statistical foundations of psychometrics now known as classical test theory. Binet was undoubtedly familiar with Spearman's contribution. In 1904, just before Binet accepted the commission to develop an intelligence test, Spearman published his paper *The proof and measurement of association between two things*. In this paper, he raised the issue of measurement reliability and showed how, by repeating independent measurements, reliability could be both increased and estimated. In the same paper, he introduced his well-known correction for attenuation for the Pearson product-moment correlation coefficient. Binet was guided by the same insights: Test items in themselves are unreliable; only by combining observations from a number of test items—thereby increasing the length of the test—can the ideal of a reliable instrument be approximated.

Another important aspect of Binet's contribution to test theory is that he developed a *standardized* test. The Binet-Simon (1905) paper features extremely precise guidelines for testing materials, administration, scoring, and interpretation. These very detailed descriptions were given with no other goal in mind than rigorous standardization. Binet hoped to develop an objective measurement procedure, in which test scores become reproducible and are independent of the personal characteristics of the tester; explicit standardization of materials and procedures was considered essential to this effort. Further, he desired not only to measure, but also to define intelligence. By standardizing the testing procedure, all disturbing factors possibly explaining test performance were eliminated, and the only remaining factor that could influence performance systematically was the ability of the persons. In other words, standardizing the test procedure operationally defines the ability it is supposed to measure.

Binet's final contribution was his notion of *test norming*. Binet calibrated his items by systematically gathering empirical data about their difficulties for various age groups. Using the data from this norming study—based on data from a small sample of children in 1905, but using data from larger samples in later revisions of the test—he also defined a mental-age scale for the test scores. The idea of using test data from well-defined populations to establish properties of test items and to

define test score scales was the key to an entirely new area of inquiry.

The "Classical Complex"

Binet's notions of test lengthening, standardization, and norming mark the beginning of classical test practice. As already noted, theoretical parallels are found in the work of Spearman, who originated classical test theory. Together, practice and theory can be described as the *classical complex* in testing. They constitute a system of basic ideas, theoretical insights, mathematical formalizations, and practical rules that has influenced test research deeply. The internal coherence of the complex has not been the only source of its longevity; there is an external reason as well. This is the analogy between its basic ideas and the paradigm of experimental control by manipulation and randomization which, at the time, made such a deep impression on the behavioral scientists as members of a young empirical discipline. Experimental effects came to be viewed as the joint result of systematic and random factors, and the same conception took form in test research. Test scores were seen no longer as the sole result of person ability, but also of random factors in the environment, in the person, or in the test procedure itself.

Spearman formalized this conception as a simple linear test score model with a true score and a measurement error. (It was no coincidence that this model was identical to an analysis-of-variance model with one random factor and that, at the same time, analysis of variance was developed as a formal model for an experimental design with manipulation and randomization.) And it was Binet who reacted to it by constructing his test starting from the principles of standardization and test lengthening. The former allowed him to increase the systematic component in his test scores; the latter reduced the influence of the remaining random effects. Moreover, standardization provided the desired objectivity, and the combination of standardization and a long test enhanced the validity of his instrument. The idea of norming tests and items on standard populations could, in turn, be inserted in the probability-based classical test model.

It was no surprise that the standardized test soon became the prototype of all measuring instruments in the social and behavioral sciences. The practical developments culminated in 1954 when the American Psychological Association published its *Technical recommendations for psychological tests and diagnostic techniques*. Previously, Gulliksen (1950), in his *Theory of mental tests*, had brought together the theoretical foundation of this complex in an impressive fashion. Many papers and textbooks propounded the advantages of standardized measurement. Those who still favored "clinical judgment" found it difficult to prevail after Meehl (1954) published the results of a study comparing the predictive powers of clinical judgment and statistical prediction in various situations of counseling, personnel management, and therapy. In no fewer than 19 out of the 20 cases investigated, personal judgments performed worse. The classical complex, with its standardized test, apparently was here to stay.

Some Practical Problems

Yet, at an early date, some hesitant objections could be heard among test researchers and users. In 1925, Thurstone wrote that he had problems with Binet's test scores and that he wanted scores with an equal measuring unit. In order to achieve this he designed, as he called it, an "absolute scaling method." Some time later, Loevinger (1947) also addressed the issue of scale properties. She indicated that for test scores to be acceptable, measurement on a scale independent from the group on which the test and items were normed was required. Comparable observations were made in passing by Gulliksen (1950). Later such observations became louder and more emphatic. Also, it became more and more clear that these objections were not merely academic but involved various practical problems. Following are some of these problems:

1. In principle, for any domain of knowledge or skill, a multitude of different test items can be devised. Thus all tests are selected from a virtually infinite domain of items. The possibly low representativeness of a test for its domain

is known as the problem of content validity. Because many item selections are possible and each has the same right to be defended as a "standard test," a serious problem arises: Classical test theory gives different true score scales for different tests. It is even possible that the same population of examinees is ranked differently by different scales. In such a case, the assignment of scores on a standard test can no longer be considered a serious attempt at measurement. The only solution to this problem would be a method for which it does not matter which selection of items is administered; in other words, a method that locates examinee performance on each possible selection of items on the same scale. Such a method would involve the entire domain of test items, and at the same time would provide a comprehensive solution to the problem of content validity.

2. Standard tests depend upon standard populations, which rarely present themselves. Differences in use of language, age groups, and curricula preclude the use of one standard instrument for all persons. Hence, different versions of a test are often required—occasionally even versions differing systematically in some property (e.g., difficulty). Still, the aim should be to locate all persons on the same knowledge or ability scale.
3. A single standard test is insufficient when the same test has to be administered twice to the same examinees. This is often the case in research projects with a longitudinal design, for instance, when effects of educational measures are to be evaluated using a pretest-posttest design. Because examinees are able to recall the content of test items, the researcher must resort to a different selection of items for the second test. But if this is done, it becomes possible to explain differences between the two test scores in terms of experimental effects as well as changes in examinees. Again, what is needed is the possibility of comparing examinee performance independent of arbitrary item selections.
4. The same problem can be met in research designs with a transversal aspect. For example, the aim of national assessment studies in education is to obtain a cross-section of the curriculum outcomes for a part of the educational system. Often a large number of curriculum elements must be covered, but the available testing time per student is restricted. If it were possible to compare performance independent of the items, for each curriculum element different items could be given to different students and an efficient design would be possible.

This list illustrates only a few practical problems; others could easily be added. Clearly, serious arguments can be made against the classical ideal of a standard test for a standard population. What appears to be needed are tests that can be composed flexibly, but still yield scores on the same scale.

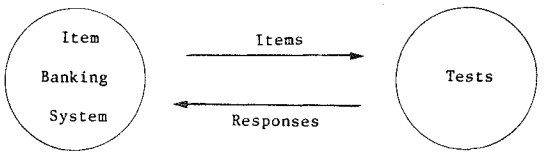
Test Item Banking

A new practice in test development is *item banking*. Two conditions have made this possible: the introduction of item response theory in psychometrics, and the large-scale introduction of the computer in modern society. This Special Issue brings together papers that provide an overview of a number of important issues and concepts of this emerging technology.

The basic idea of item banking is represented by the diagram in Figure 1. Imagine a set of test items all measuring the same domain of knowledge or ability. This set is the starting point of an item bank. At first, not much is known about the quality of the test items; at this stage, a test can only be selected using a priori estimates of item quality. As soon as a test is composed from the bank, the item responses can be fed back into it. They are used not only to score the test but also to estimate the properties of the items and to diagnose their quality. A central element in this process is the item response model, selected on the basis of plausible assumptions about the behavior of the items and with separate parameters for their properties.

The presence of feedback in an item banking system is a powerful feature. Each time the responses are fed back into the system, the estimates

Figure 1
A Simple Representation of Test Item Banking



of the item parameters can be updated and a better hold on the quality of the items is obtained. Hence tests can be composed that are in better agreement with the users' specifications. This, again, leads to more precise estimates of the item parameters; and so on.

Eventually, tests composed from the item bank fully meet their users' wishes: They may be long or short, easy or difficult, measure very accurately at a certain point of the scale or somewhat less accurately over a broader range. Also, several versions of a test with identical properties can be composed, as well as versions differing systematically in only a certain respect (e.g., difficulty). In a fully functioning item bank, responses can be retained for periodic diagnoses of item quality. Items showing a deficiency can be reformulated and provided with new parameter estimates for their next use. In the same fashion, new items can be added to the bank which, after some time, have known properties and can safely be inserted in tests.

It seems natural to computerize item banking systems. In fact, practical implementation of item banking procedures could not occur without a computer. The storage, cataloging, and retrieval of test items could be done by hand, but this soon becomes inefficient for larger banks. For the numerical aspects of item parameter estimation and updating, as well as for test scoring, a computer is absolutely necessary. Further support may consist of estimation and updating of norm distributions, administrative applications, and automated test design. This is not a full list of possible applications; for a more complete systems analysis, see the paper by van Thiel and Zwarts in this issue.

Item Response Models

Item response theory emerged in the 1950s and 1960s as a reaction to classical test theory. Unlike classical test theory, item response theory is focused not on test scores of random samples, but on the responses of individuals to individual test items. These responses are modeled as the outcome of a stochastic experiment in which the probability of a given response depends on a number of different parameters. Usually, the parameters can be classified as person and item parameters. Dependent on the content of the items, the person parameters can be interpreted as measures for the ability, the level of knowledge, or the skill of a person. The item parameters represent psychometric characteristics of the items (e.g., their difficulty). Introductions to the various item response models available are given in Birnbaum (1968), Fischer (1974), Hambleton and Swaminathan (1985), Lord (1980), Rasch (1960), Wright and Masters (1982), and Wright and Stone (1979).

In item banking, the first step is to estimate the item parameters in the model. As this can be construed as locating the items on the measurement scale, this stage is usually called the item calibration stage. In principle, two different strategies are possible here. The first is to estimate item parameters for separate tests from the bank, and then link the estimated parameters onto a common scale. The paper by Vale in this issue reports results on the accuracy of several designs for item parameter linking. The second strategy is to pool the data in advance and then estimate the parameters on a common scale. An optimal sequential procedure for this is given in van der Linden and Eggen's paper. The use of response models for item banking is not restricted to dichotomous items. Masters and Evans, in their paper, describe a model for banking test and questionnaire items scored in ordered response categories. As is clear from Hornke and Habon's paper, item response models can even be used to detect the cognitive operations needed to define homogeneous item banks.

As soon as the item parameters are known, the person parameters can be given full attention and the model can be used as a measurement model. For each next person the value of the person parameter can be estimated from his or her responses on a selection of items from the bank. In this ap-

proach, measurement takes the form of statistical estimation. The important point to note is that in the estimation equations, all item parameters have known values. Thus, the scores are automatically corrected for the properties of the items and all persons are placed on the same scale.

A "Modern Complex"

Classical test theory and the ideal of the standard test, in the decades since their advent, have become so intertwined that they can be considered as the theoretical and practical sides of the same development. Their evident successors, item response theory and the item bank, seem to exhibit a similar interdependence. Item banking without item response theory is infeasible. But it is equally true that the potential of item response theory can only be realized in combination with item banking. Both points will now be illustrated somewhat further.

An important activity in classical testing practice was the estimation and periodic updating of the norming distribution of the standard population. Although it no longer seems useful to assume sampling of persons from an exclusive standard population, the use of norming distributions for describing the relative standing of test scores with respect to relevant groups may still be a helpful device. In education, for instance, it is common to provide reports of test performance using percentile scores based on groups of students following the same curriculum, of the same age, in a certain school district, and the like.

For item banks it is infeasible to estimate norming distributions for all possible tests; the number of possibilities simply prohibits this. However, this problem can efficiently be solved using item response theory; it then becomes possible to build up distributions for relevant groups over the ability parameter in the model. Each time a test is selected from the bank, these distributions can be transformed into norming distributions over the test score. In this transformation, the item parameters again play a basic role. For a computerized system, the computations involved are easy to execute. The special advantage of the procedure is that it does not matter which items were used to build up the distributions over the ability parameter. It is not even necessary that one of these items be inserted in the tests for which norming distributions are generated. More efficient use of response data is difficult to imagine.

The use of item response theory in item banking systems also affords the possibility of automated test design. The relevant quantities for this are the test and item information functions, which describe the information in the test score and item response variables on the ability parameter as a function of its possible values. Automated test design is possible when the test user specifies, in addition to other possible constraints, a target for the test information function. In general, the shape of this target will depend on the intended use of the test. The computer then selects the items such that, subject to the other constraints, their test information function best approximates the target. A successful implementation of this strategy in an item banking system is possible only if algorithms for this selection procedure are available. A promising solution to this problem is given in Theunissen's paper.

A final contribution of item response theory to computerized item banking is the possibility of adaptive test administration (Weiss, 1982, 1985). In adaptive testing the items are not administered simultaneously but one at a time, the advantage being that the selection of the next item can be based on the responses to the previous ones. The test can thus be adjusted to the knowledge or ability level of the examinee, and a considerable saving of test length is possible. Such procedures are shown to full advantage in computer-aided instructional systems, where learning also takes place at the terminal and small numbers of items are regularly administered to monitor achievement.

The above illustrates how item banking may profit from the use of item response theory and how their combination has made possible a new testing technology. Conversely, item response theory also needs the practice of item banking to be fully applicable.

This will be illustrated using three arguments occasionally put forward to show that item response theory has only limited practical meaning. However, this misconception can arise if the usual practice of a single test administration followed by item analysis is assumed, but disappears with systematic item banking.

The first argument is that the number of responses needed to estimate test item parameters is too large for regular application. Indeed, this number may be too large for a one-time application on a small scale. In an item banking environment, on the other hand, items are permanently available, they can be inserted in tests more than once, and every response can automatically be retained for item calibration. These advantages grow if item banking takes place in a network of users. The problem is not that item response theory requires so many responses, but that in the absence of an item banking system, large numbers of responses are discarded.

The second point focuses on the issue of model fit. This certainly is a delicate point if item response theory is used in the traditional fashion, that is, for item analysis of a test administered only once. In this case, if some items show a poor fit, the original intentions for the test cannot be realized. With systematic item banking, the situation is much more favorable, because the analysis of model fit is part of the process of improving the quality of the items. New items are considered merely as first versions that require adjustment and repeated tryout until they are of satisfactory quality. Moreover, an item bank's internal structure can be based partly on the results from model fit analyses. Items not fitting together may do so in subdomains or could be the starting point for a new domain. A nice illustration of this procedure with a practical result is given in Hornke and Habon's paper.

The final argument is not fundamental and has only historical meaning. It could be argued that item response theory requires the availability of computers and sophisticated software, whereas classical analyses sometimes could reasonably be executed by hand; this would prevent it from being applied, for example, in schools. This argument does not hold for item banking. Moreover, by now computers have taken their place in almost every institution in modern society.

Conclusion

The practice of item banking and the theory of item response models are interdependent. Together they have introduced a fully new technology. Although some critical observations may be made as to its reception in computer-based instruction, such as those provided in Baker's paper, it is expected to pervade the use of tests in education and psychology for the next decades.

References

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington DC: Author.

Binet, A., & Simon, Th.A. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *l'Année Psychologie, 11,* 191–336.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

DuBois, P. H. (1970). *A history of psychological testing*. Boston: Allyn and Bacon.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* [Introduction to the theory of psychological tests]. Bern: Hans Huber.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

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

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs, 61* (Whole No. 285).

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

Meehl, P. E. (1954). *Clinical vs. statistical prediction*. Minneapolis: University of Minnesota Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedogogiske Institut.

Spearman, C. (1904). The proof and measurement of

association between two things. *American Journal of Psychology, 15,* 72–101.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16,* 433–451.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473–492.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53,* 774–789.

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

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: Mesa Press.

Author's Address

Send requests for reprints or further information to Wim J. van der Linden, Twente University, Department of Education, P.O. Box 217, 7500 AE Enschede, The Netherlands.