Robert L. Thorndike
Teachers College, Columbia University

International Research on School-Based Assessment

The five articles in this issue provide a dramatic illustration of the wide diversity of conditions under which psychologists concerned with the assessment of children undertake research and provide service. One is amazed that Dr. Saigh found it possible to gather and analyze data in strife-torn Lebanon, and his report provides testimony to the devotion with which he pursued his endeavours under most difficult conditions. In the People's Republic of China, also, political strife and turmoil have hardly provided a fertile soil for assessment research. By contrast, the United States is swarming with school psychologists, some providing service in the schools and some who are situated in universities where research and development is encouraged and supported.

Though the authors of the articles in this issue tend quite generally to affirm that assessment should be comprehensive, synthesizing information from a range of different sources, it is clear that the central concern of assessment in the educational context for most countries, and especially of assessment research, has been and will continue to be appraisal of the general level of the pupil's cognitive functioning. Assessment research on instruments or procedures can be directed at any one of three levels: (1) development of an instrument or procedure and internal analysis of its properties, (2) studies of the effectiveness of the instrument or procedure as a predictor of significant educational outcomes, or (3) studies of the usefulness of the instrument or procedure in guiding interventions to improve learning. We may consider each of these in turn, first as they apply to psychometric type instruments that yield one or more scores and then as they relate to less structured approaches such as observation and interview procedures.

The prototype of the psychometric score-yielding approach is the test of general cognitive ability, or intelligence if you will. Tests of this sort are

reported for each of the countries represented in this issue, but such instruments are probably most common in the United States. The internal analyses embrace those that are typically included in an adequate technical manual for the instrument. They include provision of adequate norms based on a representative sample of cases, evidence on the reliability of the resulting scores, and evidence to clarify the nature of the attribute or attributes that the instrument appraises. This last often involves correlational analysis of the several subscores that the instrument provides to determine to what extent all are expressions of a common underlying ability and to what extent distinct abilities are involved, and to find out how well any distinct factors confirm the theoretical constructs assumed by the authors of the instrument. It also commonly involves relating scores on the new instrument to other measures thought to appraise the same or similar attributes. Much of the literature of ability testing is of this sort.

When, as is reported by three of the contributors, a test from one national setting is adopted for use in another, questions arise as to the transferability of the test (or other instrument) to the new national setting. Where the language is the same and the cultures are as similar as is the case with Canada and the United States, the problems seem minimal, so a sensitive review by Canadian editors sufficed to convert the *Cognitive Abilities Test* which had been developed in the United States into the *Canadian Cognitive Abilities Test.* New norms are needed for the adopting country, but most of the other evidence from the country of origin on the properties of the instrument seems directly transferable. In the case of a measure of academic achievement, in addition, a careful review of the appropriateness of the content of the test to the curriculum of the receiving country seems called for.

With larger cultural differences, and especially where translation to a different language is required, a complete re-evaluation of the instrument in the adopting country would seem essential. This is brought out by LaVoie's discussion of the use of the *Wechsler Intelligence Test for Children-Revised* (WISC-R) in China, where subtest scores varied dramatically from those obtained in the United States and certain subtests appeared to have markedly low relationships to total IQ, though the test as a whole still functioned as a good predictor of academic achievement. By contrast, in Israel the complete evaluation of the Israeli adaptation of the WISC-R appears to have confirmed the appropriateness of the original test with only minimal changes.

The reasonably satisfactory effectiveness of general aptitude tests as predictors of school achievement has been repeatedly demonstrated over the years. It is confirmed by each of the authors in the present issue. Questions arise, however, as to whether a given level on the aptitude measure signifies

the same level of expected achievement in all segments of a particular society. This is the ever-present issue of test "fairness" or test "bias." Does a particular IQ in an 8-year-old, for example, forecast the same level of mathematics achievement at age 10 for black and for white children? For children from upper socioeconomic levels and those from low socioeconomic levels? For the child of recent immigrants from East Asia as for a native-born child? These are all groups that show differences in mean score on the aptitude tests, and these differences between groups describe a situation that does exist. The question is whether there are corresponding differences in other aspects of performance so that the score provides a fair forecast of the expectations for the individual's future. This is a problem that Canada is aware of in relation to its aboriginal population and to its immigrants from many parts of the world. It is one that Israel faces as immigrants from Asia and Africa are compared with those from Europe. It is one that the United States faces with its black, Hispanic, and native American groups as well as with its recent flood of immigrants from East Asia.

There is certainly no universal answer to the question of "fairness." The answer will differ for different groups and for different circumstances. Where prediction is limited to academic achievement, the evidence tends to indicate that there is no underprediction of the later academic performance of black pupils in the United States. However, for non-English-speaking groups the situation may be different. And if optimal interventions are introduced, especially at a very early age, the original expectations may need to be significantly modified. Zeidner reports that in Israel the substantial differential between Eastern and European immigrant groups is diminished in second generation Israelis, and this suggests that in time the difference might largely disappear. It is unfortunate when the existence of group differences in mean test scores leads to wrangling about the use of the tests, rather than to a constructive search for interventions that might reduce the differences, or serve to disconfirm the predictions that the test scores currently support for both the majority and minority groups.

This leads to our third type of research – research seeking to determine in what way the measures of aptitudes can guide interventions that will enhance pupil achievement. This field of research has sometimes been spoken of as "trait-treatment interaction," modifying the educational treatment to fit it to the appraised standing on the predicting trait. As tests of more specific abilities – visual, spatial, verbal, quantitative – have multiplied over the past 50 years, the hope has been expressed that particular forms and organization of instruction might be found to prove especially effective for persons with particular patterns of abilities. However, in their review of the research literature, Cronbach and Snow could find little evidence that this was in fact

the case. What evidence of interaction they did find was largely between level of general cognitive ability and form of instruction. There is some indication that higher aptitude children progress better with discovery approaches to learning while low aptitude children make better progress with relatively highly structured presentations. In addition, some research indicates that low aptitude children lack strategies for learning, and benefit from systematic instruction and practice in the use of learning strategies.

From the beginning days of testing, grouping by ability, or "streaming," was undertaken as a form of adaptation to differences in ability level, but the accumulated evidence, going back almost 70 years now, shows little indication that children learned more in homogeneous than in heterogeneous ability groups. Possibly this was because little perceptive adaptation of instruction was made to fit the characteristics of the more homogeneous group. In the final analysis, it is in instructional adaptations to individual children that result in greater learning that the justification for testing must be found, and research effort for the future should be directed to identifying such adaptations in a form that teachers can and will use.

The same three types of queries can be addressed to inventories, questionnaires, and other types of procedures designed to assess personality characteristics. What attribute or attributes does the procedure appraise and how well does it appraise them? What outcomes of concern to society can the procedure predict? And what guidance can it provide for constructive educational interventions?

Whenever an instrument or procedure yields one or more scores that are conceived to provide an index of some describable and definable attribute, the psychometric properties of the score or scores can be studied in essentially the same way that has been applied to the familiar instruments for measuring abilities. A good deal of the research reported here from the different countries is of this sort, dealing with evaluation of internal consistency, stability over time, and concurrent correlates of the several scores derived from various inventories and questionnaires. Reference is also made to a good deal of work on establishing adequate norms to which scores of individuals may be referred. As academic correlates are sought for the various measures of attitude and adjustment, the bulk of the work appears to deal with concurrent relationships. One could wish for a clearer picture of the extent to which these instruments permit a prediction of further educational progress and provide a guide to constructive intervention.

To take one specific example, reference is made to measures of "test anxiety," and some moderate negative correlations are reported between

reported anxiety and academic performance. Leaving aside for now the question of whether pupils do poorly because they are anxious or are anxious because they are doing poorly, one would like to see studies that use the anxiety score as the basis for some describable modification of the instructional situation and which follow up to determine whether measurable improvement results.

Scores and ratings of personal characteristics are presumably gathered in order to introduce constructive interventions in the education of the child. In a few instances (e.g., Gresham and Reschly as reported by Cummings and Laquerre) explicit attempts are made to relate types of adaptive behaviour deficits to the appropriate treatment. It is not clear, however, that a systematic research design has been set up to apply the recommended intervention and to follow up in order to determine how effective it has been.

It must be admitted that it would be extremely difficult to even begin to achieve rigor in research on a problem such as this. Cases presenting a certain personality deficit, for whom a particular intervention appeared to be indicated, would be likely to be relatively few in number and scattered. Providing adequate monitoring of the manner in which the specified intervention had been introduced, defining the nature of the changes in academic or personal progress that were to be looked for, and setting up procedures for appraising the extent to which the sought-for progress had been achieved is indeed a formidable undertaking. It is hardly surprising that it has seldom, if ever, been carried out.

In several of the papers in the series there is an affirmation of the view that assessment should be comprehensive and should include, and in fact rely fairly heavily on, nontest, nonquestionnaire types of procedures – such approaches as direct observations of pupil behaviour and/or interviews with parents and teachers. These more fluid approaches are felt to add important information that is not well obtained by more formal and structured approaches. The output from these procedures may often be a discursive, ideographic account of the assessor's impressions as derived from all sources of information about the child.

In the case in which observation or interview leads to some type of standard datum, as when an observer records the proportion of "time on task," or the frequency of certain defined categories of aggressive actions toward other children, or when an interviewer, or the reader of an interview protocol, rates the home for one or more attributes such as "degree of parental support of learning," that datum is in theory susceptible to analysis in much the same way as any other score. Evidence of reliability can be sought by taking

successive behaviour samples or by getting a protocol scored by more than one independent rater. Generalizability can be tested by getting behaviour samples in different settings. And the current correlations of the score with academic or other aspects of school performance can be determined. Though time-consuming to obtain and unwieldy to work with, these data can be dealt with in the same way as test scores or any other quantitative data.

But it may well be contended that the effort to generate scores from the individual and discursive synthesis that constitutes the essence of informal assessment procedures loses what is vital to the procedures and is a travesty on true clinical assessment. It may be asserted that the assessment must be dealt with directly in its complex, discursive qualitative form. To the extent that this is accepted, any attempt to do research on the procedures faces the most formidable obstacles. What is the independent variable whose impact is to be determined? Is it simply the global fact that an assessment has been carried out compared with no assessment, and that some sort of intervention was introduced? Would it be technically feasible and at the same time ethically defensible to set up a comparable control group from whom assessment had been withheld in spite of their presenting problems similar to those of the children sent for assessment, and then to compare the two groups on some index or indices of educational or personal progress? It is not to be wondered that there is a dearth of such research.

The formidable problems of conducting research on these discursive, ideographic approaches to assessment make it seem likely that there will always remain an unbridgeable gap between assessment research and assessment practice. Research will be directed primarily at those procedures that generate some type of score or scores that can be obtained for each member of a group and can be related to other facets about that individual – to other aspects of input or to aspects of outcome. Practice will, in part at least, be concerned with a nonquantitative synthesis that is unique to the particular individual and that gets its vindication from the fact that it appears to provide one or another type of benefit to the pupil or comfort to the teacher. It appears that this disparity must be recognized and accepted.

Robert L. Thorndike, Professor Emeritus, Teachers College, Columbia University, received his higher education at Wesleyan University (1931) and Columbia University (1935). He was Professor of Psychology at Teachers College, Columbia University from 1935 to 1976. He served in the U.S. Army Air Corps Aviation Psychology Program (1942-46), retiring with the rank of Major. He has served as president of the Psychometric Society, Educational Research Association, American Psychological Association (Divisions 5 and 19). Among his many publications are: *Personnel Selection* (1949); *Concepts of Over- and Underachievement* (1964); *Applied Psychometrics* (1962); co-author of *Evaluation and Measurement in Psychology and Education* (with Elizabeth Hagen), 1977; *10,000 Careers* (1959); and *The Cognitive Ability Tests* (Form 4) (1985).

Professeur émérite au Teachers College de l'Université Columbia, ***Robert L. Thorndike*** a fait ses études supérieures à l'Université Wesleyan (1931) et à Columbia (1935). Il a été professeur de psychologie au Teachers College de l'Université Columbia de 1935 à 1976. Il a servi dans l'Armée de l'air américaine dans le cadre du programme de psychologie (1934-1946) et en est sorti avec le rang de major. Il a été président de la *Psychometric Society*, de l'*Educational Research Association*, de l'*American Psychological Association* (divisions 5 et 19). Parmi ses nombreuses publications, signalons *Personnel Selection* (1949), *Concepts of Over- and Underachievement*(1964), *Applied Psychometrics* (1962); il est co-auteur de *Evaluation and Measurement in Psychology and Education* (avec Elizabeth Hagen), (1977), de *10,000 Careers* (1959) et de *The Cognitive Ability Tests* (formulaire 4) (1985).