have only limited ranges of convenience, or appropriateness: thus, the construct "printed, versus not printed" is relevant to construing books and syllabi, and also dresses and skirts, but it cannot be used meaningfully in relation to people, school subjects, or careers. Further, it is likely that a complete range of relevant constructs will be elicited only if the elements constitute a fully representative sample. The number of elements used commonly varies from around 10 to 25. The greater the number of elements, the more likelihood of representativeness being achieved, though in some circumstances, the subject matter may mean that fewer elements are available, or the nature of the sample (e.g., less able pupils) may mean that fewer are desirable.

Just as with elements, the greater the number of relevant constructs that are elicited, the greater their likelihood of being representative. Again, optimal numbers range from around 10 to 25, though most people appear to use fewer than 20 different constructs in relation to people, and some use as few as one or two. Constructs can be classified in a number of ways, such as into physical, for example, "tall" (versus the opposite), situational, for example, "is a pupil from this school" (versus the opposite), behavioural, for example, "writes quickly" (versus the opposite), and psychological, for example, "is likely to do badly under exam pressure", (versus the opposite); or into vague, for example, "is OK" (versus the opposite), and excessively precise, for example, "is a medical student" (versus the opposite).

Repertory grid data can be analysed in a wide variety of ways, and manipulated using a wide range of procedures. Thus, the interviewer may simply be concerned to note which constructs a given individual uses, and in the case of laddering or pyramid construction, also to infer relationships between constructs. Alternatively, elicited constructs can be used in idiographic or nomothetic instruments, relevant to a particular individual or group of individuals, or to particular elements. In some forms, repertory grid data can be subjected to statistical or other mathematical procedures, or used as the basis for devising interactive computer programs.

Mathematical analyses of repertory grids have been used to calculate a number of "structural" indices concerned with relationships between elements and constructs. Notable among these are cognitive complexity, cognitive differentiation, and articulation, which are measures of tendency to construe elements in multidimensional ways; identification and assimilative projection, which are concerned with perceived similarity between self and others; and constellatoriness and the coefficient of convergence, which measure similarities in the use of constructs. Mathematical techniques commonly applied to repertory grid data include cluster and factor analysis, and on the bases of these, diagrammatic representations of element and construct relationships have been devised.

The applications of repertory grid technique fall into two principal groups, "static" and "dynamic". In both groups, repertory grids can either be idiographic, in which case the individual's own constructs (or a mixture of own and provided constructs) are used, or nomothetic, in which case, for purposes of comparison, provided constructs are used commonly (though not exclusively). Examples of the "static" use of repgrid data are determination of a student's (or teacher's) perceptions of, say, self, family, and peers, or perceptions of self in relation to career opportunities in vocational guidance. "Dynamic" use of the repertory grid can involve completion of a comparable grid on two or more occasions, in order to give a measure of the extent to which an individual's construct system changes over time. This can be useful, for example, in studying the development of self-awareness, or friendship development. Alternatively, repertory grid data can form the basis of interactive computer programs, for example, in decision-making exercises.

### *Bibliography*

Fransella F, Bannister D 1977 *A Manual for Repertory Grid Technique.* Academic Press, London

Pope M L, Keen T R 1981 *Personal Construct Psychology and Education.* Academic Press, London

Shaw M L G (ed.) 1981 *Recent Advances in Personal Construct Technology.* Academic Press, London

## Diagnostic Assessment Procedures

**G. Delandshere**

A diagnostic assessment procedure is a means by which an individual profile is examined and compared against certain norms or criteria. Diagnostic assessment focuses on individuals whereas diagnostic evaluation is centered on schooling processes such as the curriculum, program, administration, and so on. In both cases the task is to determine the strengths and weaknesses of the individual or process under study. Testing is used for diagnostic assessment of student learning problems and focuses on the construction and utilization of tests as well as their interpretation. Mastery tests can also be used for diagnostic purposes since they describe the teaching/learning process and student performance. Finally, a distinction must be made between diagnostic procedures used for assessing specific learning disabilities and those used in the regular classroom for assessing learning difficulties. Only the latter will be discussed here.

Different explanatory mechanisms, or analogies, have been used in this field but their adequacy can be questioned. When the term "diagnosis" is used, it is often assumed that a medical model is implied. A set of objectives is to be reached; the student's actual performance is then measured and any pattern of discrepancies is examined and analyzed to find a "remedial treatment" or "prescription." Another analogy found in the literature is based on computer technology and refers to "procedural bugs" or "student bugs" when describing difficulties in learning. Not all researchers and educators accept the value of these analogies because most learning problems involve teaching problems as well. Hence, diagnosis should not focus on the student alone. There is nothing "pathologically" wrong with a student who has learning problems, as there is in the case of a diseased patient, but rather an incompatibility exists between a particular teaching method and the student's learning activity or cognitive style. These issues are important and need to be kept in mind when considering the uses of diagnostic testing and diagnostic evaluation in the classroom.

## 1. Diagnostic Assessment

Diagnostic assessment differs from summative types of assessment, such as minimum competency tests or final or certification examinations. The difference between summative, formative, and diagnostic assessment lies in the type of question each of them is addressing. Summative assessment is concerned with a final product whereas formative assessment provides a description of student progress. Diagnostic assessment draws a profile of student achievement, considering (a) the discrepancies between expected and actual achievement, (b) the cause for such discrepancies, and (c) appropriate "remedial treatment." This last type of assessment requires the definition of clear learning objectives and assessment techniques. A global approach to diagnostic assessment relies on teacher observations, analysis of student work, results obtained on achievement tests, and is not concerned with the construction of specific diagnostic instruments. From this perspective, diagnosis of student learning difficulties cannot be made through the examination of a single specific skill. Rather, student achievement on a broad range of objectives should be documented. Only after examining the entire student profile will it be possible to evaluate the possible causes of learning problems and make an accurate diagnosis. This "case study" approach is not intended to be used for each student in a classroom but only for those who experience persistent learning difficulties. Such a thorough examination is time consuming, and most students do not require this kind of assessment.

Effective diagnostic assessment relies on the right kind of data being collected and their correct interpretation. The danger exists that a diagnostician might define a limited set of learning-problem categories, label students according to these categories, and subsequently only prescribe familiar or readily available treatments. At each level of this process the validity of the diagnosis and the "treatment" is in question. Continuous evaluation of "prescribed treatments" is necessary. In addition, empirical data are needed to test the efficacy of the model as a whole (Thomas 1981).

## 2. Diagnostic Testing

If diagnostic testing is defined as providing feedback to teachers and students regarding their strengths and weaknesses, almost any test would be diagnostic. In effect, a score on any standardized test gives some indication of the student's performance and also informs the teacher as to how successful he or she was in teaching the material. But the total score does not give any real information concerning specific areas of difficulty or their causes. Examination of the individual items would tell us which answers are correct and which are wrong. However, most available tests have not been constructed for diagnostic purposes; they have too few items per objective and the analysis of the wrong answers cannot lead to a diagnosis of learning problems. Diagnostic testing is seen as one component of the ongoing teaching/learning process. In order to use tests for diagnosis, they should be specifically designed for that purpose.

Guidelines for constructing "good" multiple-choice tests for diagnostic purposes are needed. As with any other test, a clear definition of learning objectives is required; but each test addresses a very specific skill (e.g., addition, subtraction, etc.) and the items are designed to test a particular subskill from various perspectives. Each item provides information to the student and the teacher since each distractor has a precise meaning in terms of learning difficulties and possible remedial strategies. Here the emphasis is on the significance of each particular answer and the response patterns in general. The total score has no real importance or meaning, and guessing is not usually a concern in diagnostic testing since evidence that a student guesses does not provide the specific information needed for diagnosis and remediation.

A consensus does not exist concerning the use of distractors as sources of information for diagnosis. The three main approaches to this issue can be described as follows. One view maintains that the choice of a given distractor has a specific meaning in terms of learning difficulties (Baker and Herman 1983). Working from open-ended questions, Tatsuoka and Tatsuoka (1983) analyze the responses to discover which rules are followed in finding the answers. And finally, in adaptive or sequential testing, the only information taken into consideration is whether or not the answer is correct. The last approach does not analyze or ascribe any meaning to the answers given by students.

In adaptive or sequential testing, the items are selected on the basis of the student's responses to previous items. More specifically, a correct response leads to a more difficult item, whereas an incorrect response leads to an easier one. Item selection does not depend on which distractor was chosen but in whether or not the answer was correct. This type of testing is usually administered by a computer which records the answers and selects the items. The pool of items used in one sequence is pretested for unidimensionality, and item difficulty is the basis for their selection. Finally, hypotheses about learning difficulties are formulated on the basis of the response pattern.

One problem with sequential or adaptive testing is that decisions are often made based upon the answer to a single item. However, certain diagnostic characteristics have been attributed to these tests. As they are currently designed, their main advantage is in improving the precision of measurement since each individual receives a different set of items corresponding to his or her level of performance. In order to effectively use adaptive testing for diagnostic purposes the number of items on which diagnostic decisions are made would need to be increased. Also, since adaptive testing has the flexibility to readily adjust to the subject's performance level, potential exists for its use in diagnostic assessment.

A final research perspective found in the literature attempts to explain why errors are made by identifying the "misconceptions" or "bugs" that produce them. The first task is to construct a "procedural network" for each skill under study by breaking down problems into the requisite subskills. What are the correct and incorrect procedures that can be followed in attempting to perform a task? Breaking the skill down into subskills is crucial because, in order to be efficient, a diagnostic model "must contain all of the knowledge that can possibly be misunderstood by the student or else some student misconceptions will be beyond the diagnostic capabilities of the system" (Brown and Burton 1978). Secondly, a set of items must be generated that provides opportunities for students to demonstrate all the identified subskills. Student answers are then examined to reveal those subskills which have not yet been answered. Based on these student "misconceptions" a second set of problems is administered to confirm the initial hypothesis. The information gained from this testing should suggest possible remedial strategies.

This approach has been implemented through the use of the "Buggy system" (Brown and Burton 1978), a computerized game used mostly for teacher training. The computer plays the role of the student, while the teacher has to recognize the source of student error (or bugs) and be able to replicate them. Hence, teachers become more sensitive to the causes of student learning problems. It seems evident that by following this model, single "bugs" can be easily diagnosed. However, when a student has "multiple bugs" or when different "bugs" interact, the diagnosis becomes more problematic. Also, if subskills or possible misconceptions have been omitted from the "procedural network" some of the response patterns may be interpreted as random error.

This research perspective is primarily concerned with developing a model to account for hundreds of "bugs" that can occur when skills being studied become more complex. A probabilistic model seems to be more appropriate for detecting aberrant response patterns when several hundred "bugs" are possible. Brown and Burton's model showed that it was possible to give the correct answer even when using "erroneous rules." Tatsuoka and Tatsuoka (1983) refined that aspect of the model by introducing the Individual Consistency Index (ICI). A low total score and a high Individual Consistency Index indicate that the student is using the wrong rules to solve the problems. This index can be used as a signal to point out students who need remediation and more refined diagnoses. Also, using item response theory, they demonstrate that when the possibility exists to respond correctly by using the wrong rule, the data set obtained reveals multidimensionality. In other words, the test measures different dimensions depending on whether the student masters the skill or not. This finding has definite implications not only for diagnostic testing but for testing in general and points out the importance of checking for construct validity.

All "bugs" do not affect the learning process in the same way but a typology of "bugs" or misconceptions still has to be developed. Furthermore, teaching strategies to remediate these diagnosed learning problems are not always available. Before making any concluding comments, the testing techniques of mastery learning will be briefly examined to demonstrate how they may also be classified as diagnostic instruments.

## 3. *The Mastery Learning Approach*

This model uses testing in a way that is not always considered "diagnostic," but in fact does serve this function. It involves the use of "formative" or "mastery" tests at frequent intervals to test student performance on each item as compared to a "mastery standard." Examination of these answer patterns provides information on the level of learning and indicates whether a student needs more practice or another teaching method. Some theorists in the area of mastery learning contend that the construction of formative tests is based on a theory of learning. An underlying theory is useful in designing the test in such a way as to identify not only the learning problems but also the cause of the problems.

The difference between mastery and diagnostic testing is not always obvious. Since the term "diagnostic" is used in so many different ways, confusion regarding the application of this term persists. For example, the model of diagnostic assessment developed by the Scottish Council for Research in Education (Black and Dockrell 1980) differs little from a mastery learning model. The former model uses criterion-referenced tests constructed according to a taxonomy of objectives. These objectives, or intended outcomes, are directly linked to the curriculum. While the model is designed to assess mastery it does not necessarily provide information for detecting potential sources of learning difficulties. Therefore, although this has been labeled a diagnostic model, it does not contain many of the diagnostic features of some of the models presented here.

Formative or mastery tests have some of the same objectives as diagnostic tests. One difference may be with the way they are constructed. As indicated above, mastery tests seem to correspond to a theory of learning whereas most diagnostic instruments are constructed to test hypothesized erroneous rules that students follow in solving problems. Mastery tests are attached to a teaching method; they are used to assess student learning and provide students and teachers with feedback regarding skill mastery. This method may not be sufficient for all students, and more refined instruments with better interpretation techniques may be needed to diagnose skill mastery for some students. These are not necessarily included in mastery learning programs. Generally, the type of information sought determines the appropriate kind of test to be used. If the teacher wants to know who has mastered a particular skill and who has not, almost any testing instrument will suffice. However, if partial knowledge is considered important and if it is believed that knowing in detail what kind of error the student makes helps adjust teaching methods, then multiple testing and more refined instruments are necessary.

### 4. Conclusion

As has been seen, diagnostic assessment can be conducted in a variety of ways, ranging from global evaluation to more refined diagnosis of very specific skills. The particular demands of each situation will determine which type of assessment is most appropriate. Other factors such as time and cost effectiveness will influence the decision as well. For example, some learning difficulties are a function of developmental lag, therefore intensive diagnosis would be unnecessary and unproductive.

Diagnosis of temporary learning problems can also lead to persistent "labeling" that endures even after the learning problem has disappeared. Research has shown that teachers' expectation can influence their attitudes and behavior toward students.

A final point concerns the assumptions upon which diagnostic testing rests. The first is that learning can be "decomposed" into a set of discrete subunits or subskills. The second is that items can be generated to measure accurately and validly these subskills. All the diagnostic testing models described above are predicated on these assumptions. The practical or educational value of these models, therefore, depends on the validity of these assumptions, which in many areas of education remains to be established.

### Bibliography

Baker E L, Herman J 1983. Task structure design: Beyond linkage. *J. Educ. Meas.* 20: 149–64

Black H D, Dockrell W B 1980 *Diagnostic Assessment in Secondary Schools: A Teacher's Handbook.* Scottish Council for Research in Education, Edinburgh

Brown J S, Burton R R 1978 Diagnostic models for procedural bugs in basic mathematics skills. *Cognit. Sci.* 2: 155–92

Tatsuoka K K, Tatsuoka M M 1983. Spotting erroneous rules of operation by the individual consistency index. *J. Educ. Meas.* 20: 221–30

Thomas R M 1981 A model of diagnostic evaluation. In: Lewy A, Nevo D (eds.) 1981 *Evaluation Roles in Education.* Gordon and Breach, London

## Confidence Marking

**D. A. Leclercq**

According to Ebel (1965), "the term confidence weighting refers to a special mode of responding to test . . . items, and a special mode of scoring those responses . . . . The examinee is asked to indicate not only what he believes to be the correct answer to a question, but also how certain he is of the correctness of his answer. When his answers are scored he receives more credit for a correct answer given confidently than for one given diffidently. But the penalty for an incorrect answer given confidently is heavy enough to discourage unwarranted pretense of confidence."

This article considers how the credits and penalties are assessed.

### 1. The Underlying Models

The choice of a level of confidence (from amongst those available) must be considered in the perspective of decision theory. Technical problems such as validity, reliability, and acuity of confidence answers must be based on a sound model of mental activity, on carefully written instructions, and on the selection of appropriate tariffs. The tariffs are the points attributed: for a correct response ($T_c$), for an incorrect one ($T_i$), or for an omission ($T_o$). A set of three tariffs ($T_c$, $T_i$, and $T_o$) is called a *t*-scale. The best known *t*-scale is the *St*-scale (simple *t*-scale) where $T_c = +1$, $T_i = 0$, and $T_o = 0$. The expected score on a given question (ESQ) is computed according to the following formula:

$$\text{ESQ} = (p \cdot T_c) + (q \cdot T_i) + (r \cdot T_o)$$

where $p$ = probability of a correct answer, $q$ = probability of an incorrect answer, and $r$ = probability of an omission.