DOCUMENT RESUME

ED 334 829 FL 019 275

AUTHOR Griffin, Patrick
TITLE Profiles; Validity Issues in Assessment and Reporting.
PUB DATE Apr 91
NOTE 10p.
PUB TYPE Information Analyses (070) -- Viewpoints (Opinion/Position Papers, Essays, etc.) (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Accountability; Elementary Secondary Education; *Evaluation Criteria; Foreign Countries; *Information Dissemination; *Literacy; *Outcomes of Education; *Student Evaluation; *Test Validity
IDENTIFIERS *Australia

ABSTRACT

There is a need to improve teachers' ability to process and communicate information about student learning. In Australia, literacy is an important area in which to assess educational effectiveness. The community will continue to demand evidence of the schools' success in this area. Educators need to define the literacy goals, standards, and levels to be achieved, and implement procedures to assess student progress toward them. The assessment should measure the effect of curriculum and instructional changes on student progress, and should be highly credible and endorsed by all constituencies. Measurement of discrete skills must be complemented by assessment of underlying development or synthesis. Performance assessment measures, as an alternative to standardized tests, have several drawbacks that interfere with their use for purposes of accountability. The literacy profile is a reporting system that comes close to meeting the need for accountability, but lacks a degree of external validity. Standardized tests can been combined with the Victorian (Australia) literacy profile to strengthen its validity. A similar strategy has been used successfully in the United Kingdom. (MSE)

Profiles;
Validity Issues in Assessment and Reporting.

Patrick Griffin.

Phillip Institute of Technology.

April, 1991.

There is a growing body of evidence that indicates how teacher judgements influence decisions in the interactive Phase of Teaching (McNair, 1978, Hoge and Colardarci, 1989). The teachers main consideration during reading instruction, for example, appears to be reading achievement. Teachers pace whole class instruction on the basis of whether an identifiable core group of students understand what is being presented (Clark and Peterson, 1986). Questioning is routinely used in the evaluation of pupil comprehension, learning, thinking, knowledge acquisition or task performance (Colter, 1984). Teacher decision-making, particularly in an interactive classroom context, is influenced by judgements about student learning.

The issue is whether the teacher judgement is accurate. The implications are important when judgements inform decisions regarding students for feedback, reporting to parents and other stakeholders (Elliott, Gresham, Freeman and McCloskey, 1988). Teacher judgements provide the primary data for most classroom decisions. However, it is widely assumed that teachers are generally poor judges of student attributes and that this is due to lack of perception, bias and error. Assessment schedules based on judgements are often described as informal assessment inventories. Even the name suggests that no real value can be ascribed to the judgements. Formal assessments are usually restricted to standardised, test based exercises which are set external to the classroom. "Directly or indirectly, the accuracy of teacher's assessments of student ability is often an issue in educational research. It is commonly argued that tests provide teachers with valuable information about the abilities and deficiencies of their students, from which it follows that teachers who rate their students without such information will be in error." (Egan and Archer, 1988, p.25)

Hoge and Colardarci (1989) however, after a meta analysis of a range of studies concluded that teachers differ in how accurately they judge their students' achievement. There was a generally high level of agreement between judgemental measures and standardised achievement test scores. A median correlation of 0.62 was found. Griffin (1989) replicated this level of agreement when teachers were provided with a descriptive criterion scale on which to have the judgement of achievement of reading and writing development. The data from research studies seems to support the concurrent validity of teacher judgements of academic achievement. Studies of both convergent and concurrent validity have reported consistently higher correlations than those for psychological tests, (Hoge and Colardarci, 1989), despite wide variations in methodology across studies. However there are still some unresolved issues.

Judgemental assessments do not always make it clear as to what aspect of student performance is being assessed. The work of Griffin (1990) and of Farr and Farr (1990) offer some guide in that both provide descriptive scales as a frame of reference for the judgements. Both illustrate how the judgement is closely allied to the teaching and learning process. The descriptions in the scales of increasing proficiency circumvent the issue of the validity of standardised test data. Moreover, problems associated with global judgements (high, low, etc.) are avoided or at least contracted by reference to the descriptive scales. Coladarci's (1986) criticism of a teacher's lack of explicit criteria is also avoided because of the disclosure about the teacher's specific knowledge of what the student has and has not mastered in some domain. In the case of the Griffin (1990) and Farr and Farr (1991) projects, the domains are Reading and Writing respectively are provided. As yet, however, too little research has been done on convergent and discriminant validity of judgements based on criterion scales.

In order to achieve maximum benefits from this form of assessment/teaching interface however, teachers may need to be sensitised to the extent and importance of the assessment role in the teaching process (Hoge, 1983 and Hoge and Cudmore, 1986). Experience with basic principles of measurement and assessment instruments and other devices, including norm referenced tests, observational procedures, and judgemental scales are needed. Rating and judgement scales need to be developed in line with Glaser's criterion referenced interpretation. Using this experience and improved assessment technology, there is a need to enhance teachers' abilities at analysing and diagnosing learning in children. All of these should be enhanced as a means of providing a wide range of information types suitable for a wide range of audiences.

A great deal of importance is attached to the judgements of teachers in both the teaching process and in communications with stakeholders in the education process. Reporting should emerge as a major focus of attention in education. Much of the debate related to assessment may be based more on the type of information and credibility (or validity) attached to the communication of assessment information. In an information age, educators should be looking at their methods of processing and communicating information to the variety of audiences and stakeholders in the community. The lag in development of appropriate communication protocols may prove costly in the short term as accountability pressures build.

Accountability:

The high school graduation group for the year 2000 is now in grade three. The cohort of students who will become the basis of higher education and the workforce for the twenty first century is already in school. Tertiary graduates of the twenty first century are already in secondary school. The current emphasis in industry of flexibility, adaptability and literacy underlines the importance of the school system in developing higher levels of literacy in order to make the society, the workforce of the next century in Australia more productive and competitive. It is also important that the school system recognises the broadening of the audience base for information about student learning and the increasing need for different types of data. One of the most important areas to detail information will be in monitoring literacy developments. In an information age, literacy will become the foundation stone of the social, economic, industrial and educational progress. The community will continue to demand evidence of the schools' success in this area.

Numerous initiatives have been undertaken within Australia in an effort to achieve higher standards of literacy. What was satisfactory in levels of literacy twenty years ago are not satisfactory now and certainly will not be for the beginning of the next century. The major issue appears to be not whether higher standards of literacy should be set or whether those higher standards should be assessed but rather how the literacy standards can be established. The issue then becomes how a student can progress towards them and how the attainment of those standards be determined. It is the responsibility of people in education to answer both questions. If they do not, then those responsible for the workforce and the economy will do so, because they have a strong interest in seeing literacy standards improve.

Educators need to define the literacy goals, standards and levels to be achieved and then implement procedures to assess student progress towards those goals. In order to convince representatives of the economy and industry that the standards, the procedures and the monitoring devices are appropriate, the schools and those responsible for the schools need to be accountable to the wider community in a far more open way than they have been to this point. If schools and those responsible for schools are to be accountable for achieving these required levels of literacy, the development of new accountability methods are required. The new methods will need to have certain characteristics. For example,

1. The standards need to be defined in terms that are operational. That is they should be concrete, measurable and results oriented.

2. Educators need assessment procedures that will document whether students are making satisfactory and sustained progress towards those standards or goals.

3. When sustained progress is not being made by all students, or if the progress is considered to be insufficient, measures that are able to identify those areas of schooling needing to be strengthened or modified are also needed.

4. The measures should be able to assess the effect of curriculum and instructional changes on student progress.

5. The measures should be highly credible and widely endorsed by all constituencies so that the results will receive support in the widest possible sense.

These five characteristics of an accountability system require two basic types of validity. Defining the goals, developing assessment procedures, monitoring progress and identifying effects of instruction all relate to the design of instruction and assessment. There must be a link between the processes of instruction and assessment. That is, there must be demonstrable evidence of internal validity of both the teaching and learning and the assessment and reporting processes.

The last characteristic refers to a need for the widest possible endorsement of both the instruction process and the assessment information. Both should have demonstrable and credible appeal to those outside the educational institutions but who are still part of the educational constituency that links literacy with life skills. The assessment should be reported in a manner that is convincing to the consumers of the education system and that is clearly able to be generalised to life skills and understood by specific audiences. This is another way of saying that the process, the assessment and the means of communication about teaching and learning must have external validity. Keeping the teaching and learning system within the closed triad of teacher, student and parent may be very satisfying to teachers because they believe the validity of their process is intuitively, professionally and theoretically sound. It will have internal validity but much of the value of this is lost if the information gained can not be generalised to other audiences in a manner that is understandable, usable and credible. External validity is an issue that teachers must come to grips with. Reporting to the community needs to demonstrate each of the characteristics of the accountability system outlined above. Internal and external validity are both essential in order to achieve this. However the first requirement is to establish a common meaning for the focus of attention - literacy.

Literacy.

There are numerous definitions of literacy, each with its own purpose. What needs to be avoided is the kind of omnibus definition that is becoming more and more popular in Australia that includes reading, writing, speaking, listening, critical thinking, numeracy and problem solving. These omnibus definitions of literacy as exemplified in the government green paper (Dawkins, 1989) border on definitions of intelligence which were prominent in the sixties and led to a psychometric "gold rush" seeking measures of the elusive 'g' factor. As the definitions become closer to those previously used to operationalise intelligence a more subtle danger emerges. The debate over whether intelligence can be taught will translate to whether the omnibus type of literacy can be taught.

Simple definitions of literacy may be easier to operationalise. Applebee and others (1987) defined literacy as the ability to read and write and to reason effectively about what one reads and writes. The definition is useful because it can be expanded into two particular operational skills. In order to be literate one must develop the ability to understand a variety of increasingly more difficult materials at least at a surface level and a reader should be able to analyse, evaluate and extend ideas that are presented. Such systematic reasoning about what they read and write give literate people the kinds of mastery of the written word required by more and more activities in today's society. (Applebee, et. al. 1987 p.9.) Mastery of the written word can relate literacy to life skills and set the basis for external validity of assessment and the means of communicating the results of those assessments.

The assessment instruments and the means of communicating the assessment results need to assess both the cognitive task and be sensitive to developmental changes in student performance over time. More importantly, the assessment and reporting procedures should be applicable at the individual student level, at the class, school, region and system level. It is unfortunate that many assessment instruments and reporting protocols available to schools, are unable to meet these accountability needs. Many traditional standardised (including criterion referenced) tests may have value for some assessment functions but they may not meet the accountability requirements at school, region and state levels. This is particularly true when no scale of development underpins the tests. Cannell (1987, 1989) outlined three particularly serious limitations of the use of norm referenced scores in relation to setting literacy goals and assessing student progress. Norm referenced score interpretations...

1. are unable to describe what students are able to do.
2. Can not be used to adequately assess student growth.
3. Do not maintain constant meaning over time.

Normative test scores therefore have no functional meaning. Expressing student performance in terms of other normative scales, such as grade equivalence, percentile rates or normal curve equivalence does not resolve the inadequacy of normative scores for the description of student performance in functional terms.

A normative score may be used to compare an individual's performance with that of a group at a particular point in time. However, although a student can improve over time, the amount of improvement can not be determined on the basis of norm referenced information. The most telling of the criticisms of norm reference scores is the change of meaning over time. If a standardised instrument is normed at one particular point in time, changes in student ability as well as changing the curriculum may mean that the test items and the norms no longer reflect the reality of the school system. Test publishers need to revise and re-norm published tests in order to get revised estimates and to better reflect the curriculum in student ability. However scores based on the new norms may not be on the same scale as the norm scores from the old scales. In other words the level of achievement associated with a particular percentile rank will not remain constant across two sets of norms. This makes the assessment of growth more difficult.

Assessment instruments which focus on distinct objectives related to instruction may not provide an acceptable alternative. In reading, for example, understanding initial consonant blends, vocabulary, reference or recognising the main idea may be discreet objectives whose acquisition can be assessed using objective reference tests. However instructional objectives will differ from school to school and from grade to grade. Identifying discreet objectives represents the belief that, if students can master these skill-based objectives, they will acquire the ability to understand and reason with text. Objective referenced assessment instruments are therefore valuable for instructional purposes, but they may not be all that valuable on their own for communicating to a wider audience. Exceptions to this are tests in which the objectives form a cohesive set which define an underlying scale of development.

Performance Assessment.

Of late, there have been numerous discussions about performance assessment as an alternative to standardised measures. See for example Wiggins (1989, 1990). Samples, portfolios, reports and performances are given as examples of performance assessment. The major value of these assessments is that the task is authentic in terms of the cognitive tasks that the students are required to perform. For accountability purposes however, the most important issue may be the subjective nature of the evaluation of the student performance and the manner in which this subject of assessment is reported. The same criticism may be levelled at the literacy profiles (Victoria, 1990). The profiles have strong internal validity and suit the classroom purpose well. They also offer a powerful form of reporting particularly to parents and children However if the assessment instrument is restricted to the teachers judgement, the external validity may suffer because of the subjectivity of the process. Public acceptance of assessment data requires at a minimum, the appearance of an independent external audit of the performance. This external audit of the performance may be the major requirement for external validity of school based assessment. If this is resolved then the literacy profiles may fulfil the requirements of te accountability system outlined above.

Two other issues that need to be resolved before performance assessment can be used for accountability purposes need to be addressed. These are aggregation of student performance and the assessment of individual growth and progress over time. Because of difficulties these areas, performance assessment on its own, is unlikely to fulfil accountability requirements. Given weaknesses in the kinds of reporting systems that are available, it is necessary to outline what the requirements of an assessment and reporting system should be.

Reporting on the effectiveness of an education system requires several properties of the assessment. It should:

1. Have accepted outcomes that chart student capabilities that are understood and credible to the general community. The literacy profiles can provide this basis for

reporting.

2. Show how student growth progresses over set time periods, for example a school year. The literacy profiles, with nine levels, should have sufficient sensitivity for this purpose particularly, if partial progress is recorded and reported at each level.

3. Show how change in student performance relative to standards and expectations. This requires the definition of external criteria as standards to be established.

4. Be sensitive to the impact of variation in resources, programs and instructional processes on student achievement. The profiles have already been shown to be sensitive for this purpose in the 100 Schools Study (Rowe, 1989).

This analysis indicates that the literacy profiles provide a reporting framework that can have both external and internal validity. The literacy profiles can be expected to fulfil these expectations, provided that judgements remain holistic rather than checklist oriented. The levels were neither developed nor used in initial trials as checklists. The Profiles Handbook (Victoria, 1989) presents the profiles as a series of checklists. This has had obvious benefits for teachers in the teaching and learning context, but changes the way they use the scales and hence in the way the scales relate to external criteria. The assessment should provide information in at least three areas.

1. It should be able to indicate how well an individual student, class or school has done in relation to the community's expectation.

2. It should relate the progress that a student has made since the last assessment.

3. It should be ale to show whether this amount of progress is sufficient relative to current requirements and community agreed standards.

Under these circumstances The literacy profiles as a reporting system go close to meeting the needs of an accountability system outlined above. Its weakness is in its external validity if the assessment is based solely on the teachers subjective judgement. There is then, a dilemma. It is the teachers judgement that strengthens the internal validity of the assessment, but the use of more objective measures is generally regarded by teachers as a threat to validity or a source of invalidity. The solution may lie in a linkage of the objective measures for external validity, and the teacher's judgement for internal validity. These can both be expressed via the reporting framework provided in the literacy profiles.

There is clearly a need for a new technology in assessment and reporting. Traditionally tests have been called upon to support many additional functions other than those or which they were initially designed. Reports of test results have often then been corrupted by the pressure to translate performances and test scores into generally understood terms.

In many cases the media reduces the information to report on single test items and allows generalisations to be made from this limited information. Despite the almost total lack of external validity, this form of reporting has become predominant. The media steals the initiative from educators. The initiative should be recaptured by those best able to define development in terms understandable by a range of audiences.The literacy profiles provide a common basis for the assessment technology and the reporting mechanism. This can perhaps be exemplified in the writing scale.

Writing assessment has recently moved towards the use of holistic outcome measures embodied in the writing sample. Moreover the use of writing samples, despite limitations, tends to exert a positive input on the nature of classroom instruction in writing. There are several examples of test instruments and assessment protocols that allow for the requirements of the assessment method outlined above.

The literacy profiles coupled with the TORCH test were shown to produce a composite assessment system that enabled both the TORCH test performances to be mapped on to the literacy profile scales and vice versa (Griffin, 1989). The purpose of that particular study was to show that the literacy profile scales could be used as a central reporting system and that both judgemental assessments and standardised test assessments could be mapped onto those central scales.

Farr and Farr (1991) have developed a similar system of standardised writing tasks based on graded reading material to map on to a central scoring protocol. The system is called the Language Arts Performance Assessment system which in effect assesses both reading and writing but reports only on the level of writing competence on a 4 point scale linked to 7 graded levels of reading difficulty.

Standard scripts will soon be available for Australian student writing. It is in effect a complete writing instruction package that links the teaching and learning of writing with the assessment, and enables standard reference scripts to be consulted for each level on the scale. It is possible to link the Language Arts Performance Assessment system with the Victorian Profiles in the same way that the TORCH test was linked in the past. The scales provide a description of proficiency which can be used for reporting purposes. They are not the assessment system or the assessment instrument in and of themselves.

A second integrated assessment and reporting system that is available was developed in the United States, and to some extent is available with limited Australian data, is the reading assessment package called Degrees of Reading Power (DRP) (College Board, 1986). The Degrees of Reading Power is a series of cloze tests using graded reading material of known reading difficulty or readability based on Bormouth's (1985), readability formula. The Degrees of Reading Power, unlike the TORCH test, does not use re-telling. The Degrees of Reading Power requires students to construct meaning from the prose as the prose is read. The advanced level DRP tests require the students to reason with the prose material. It is a series of multiple choice tests or more correctly it is a series of graded multiple choice tests. It differs somewhat from other multiple choice tests which typically are based on the proposition that the reading skill and the reading process can be broken into discrete components.

Multiple choice reading test items that follow a reading passage may well incorporate the view that reading can be broken down if they purport to form a reading comprehension test. In such tests the items tend to focus on discreet components or skills thought to be related to the process of reading. Alternatively they may ask the student to recall factual knowledge contained in the passage and in some cases can be answered by the student without even reading the accompanying passage.

The DRP differs from this kind of test. The items do not reflect discrete component view of reading. The test is non-diagnostic. It is a test of ability in the reading process. Like the TORCH test, the series of graded assessment passages are reported on a common scale. Unlike the TORCH test however the scores are not translated into discrete skills in a hierarchy of skills. The Degrees of Reading Power series of tests map onto a common scale that is interpreted in terms of the difficulty of the text material. The generalisation available from the Degrees of Reading Power is not made in terms of a test score but instead it is made in terms of the kinds of reading material that a particular performance might predict. That is to say a score of 50 points on the Degrees of Reading Power scale would indicate that the reader can cope adequately with a particular kind of reading material of specific reading difficulty. The strength of such an interpretation is that it is immediately understood by those outside the education system. The Degrees of Reading Power have also been mapped onto the Victorian Literacy Profiles.

So three standardised assessment systems have been mapped on to the Victorian Literacy Profiles: the Integrated Assessment System, the TORCH Test and the Degrees of Reading Power. Each can be translated to a level on the Literacy Profile Scales. Each can be shown to be generalisable to skills beyond the classroom via the literacy profiles. Each can be aggregated at class, school, region and state levels and aggregate scores and distributions can be mapped on to the Literacy Profile Scales. The linkage of the two (The Assessment Instrument and the Reporting System) provides a means by which the teacher may be given a range of standardised assessment tasks which can map on to a centralised reporting system. The standardised assessment task data can be aggregated and the aggregated data translated onto centralised reporting scales.

Without the long, expensive and difficult task of developing the standard assessment tasks, research on existing pools of assessment instruments can and should be extended in order to show how these map onto the centralised reporting scale.

In the United Kingdom the targets and levels of the national curriculum assessment scales in the language area, bear a remarkable resemblance to the Victorian Literacy Profile Scales. In the UK extensive work is being done to assist teachers in conducting assessments to map onto those scales. Considerable work has been done in showing teachers how to report at an individual level. The National Curriculum Council is developing means of testing students in order to ascertain the proportions of students that have reached particular levels within each attainment target area. These parallel the bands in the literacy profiles. A similar reporting framework can be developed for those scales.

The same process can apply in Victoria. Assessment instruments of known properties can be mapped onto the Victoria Literacy Profiles which can then be reported to the community in terms of the kinds of reading capacity that individuals have. Such a system of assessment provides information about student capabilities in the area of reading and writing; about student growth or progress over relevant time periods and about student performance relative to expected standards for particular grade levels. The combination of the profiles with their 9 levels and a series of standard assessment tasks should provide an assessment system which fulfils the requirements of accountability and enables both internal and external validity of assessment and reporting to be maintained.