

The role of measurement and evaluation in education policy

Edited by Frances M. Ottobre

Educational studies and documents 69

TD/TNC
62-894

UNESCO Publishing

EDUCATIONAL STUDIES AND DOCUMENTS

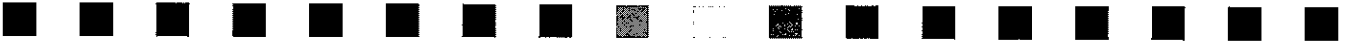
List of titles*

1. Education in the Arab region viewed from the 1970 Marrakesh Conference (E/F)
2. Agriculture and general education (E/F)
3. Teachers and educational policy (E/F)
4. Comparative study of secondary school building costs (E/F/S)
5. Literacy for working: functional literacy in rural Tanzania (E/F)
6. Rights and responsibilities of youth (E/F/S/R)
7. Growth and change: perspectives of education in Asia (E/F)
8. Sports facilities for schools in developing countries (E/F)
9. Possibilities and limitations of functional literacy: the Iranian experiment (E/F)
10. Functional literacy in Mali: training for development (E/F)
11. Anthropology and language science in educational development (E/F/S)
12. Towards a conceptual model of life-long education (E/F)
13. Curriculum planning and some current health problems (E/F)
14. ALSIED directory of specialists and research institutions (bilingual E/F)
15. MOBREAL The Brazilian Adult Literacy Experiment (E/F/S)
16. School furniture development: an evaluation (E/F/S)
17. An experience-centred curriculum: exercises in perception, communication and action (E/F/S)
18. Nutrition education curricula: relevance, design and the problem of change (E/F/S)
19. World survey of pre-school education (E/F/S)
20. The operational seminar: a pioneering method of training for development (E/F/S)
21. The aspirations of young migrant workers in western Europe (E/F)
22. Guide for the conversion of school libraries into media centres (E/F)
23. Youth institutions and services: present state and development (E/F/S)
24. Group techniques in education (E/F/S)
25. Education in Africa in the light of the Lagos Conference (1976) (E/F/A)
26. Buildings for school and community use: five case studies (E/F/S)
27. The education of migrant workers and their families (E/F/S/A)
28. Population education: a contemporary concern (E/F/S/A)
29. Experiments in popular education in Portugal 1974-1976 (E/F/S)
30. Techniques for improving educational radio programmes (E/F/S)
31. Methods and techniques in post-secondary education (E/F)
32. National languages and teacher training in Africa (I) (E/F)
33. Educational systems regulation: a methodological guide (E/F/S/A)
34. The child and play: theoretical approaches and teaching applications (E/F/S)
35. Non-formal education and education policy in Ghana and Senegal (E/F)
36. Education in the Arab States in the light of the Abu Dhabi Conference 1977 (E/F/A)
37. The child's first learning environment selected readings in home economics (E/F/S)
38. Education in Asia and Oceania: a challenge for the 1980s (E/F/R)
39. Self-management in educational systems (E/F/S)
40. Impact of educational television on young children (E/F/S)
41. World problems in the classroom (E/F/S)
42. Literacy and illiteracy (E/F/S/A)
43. The training of teacher educators (E/F/S/A)
44. Recognition of studies and competence: implementation of conventions drawn up under the aegis of UNESCO; nature and role of national bodies (E/F)
45. The outflow of professionals with higher education from and among States Parties to the Regional Convention on the Recognition of Studies, Diplomas and Degrees in Higher Education in Latin America and the Caribbean (E/S)
47. National languages and teacher training in Africa (II) (E/F)
48. Technical and economic criteria for media selection and planning in educational institutions (E/F)
49. Reflections on the future development of education (E/F)
50. Education in Africa in the light of the Harare Conference (1982) (E/F)
51. Post-literacy training and endogenous development (E/F)
52. Senior educational personnel: new functions and training (Vol. I: Overview) (E/F)
53. The articulation of school education and out-of-school training (E/F)
54. National languages and teacher training in Africa (III) (E/F)
55. Senior educational personnel: new functions and training (Vol. II) (E/F)
56. Innovations for large classes: a guide for teachers and administrators (E)
57. La discrimination et les droits de l'homme dans les matériels didactiques: guide méthodologique (F)
58. A new meaning for education: looking at the Europe Region (E/F)
59. In search of new approaches to basic education and the evaluation of learning achievement in the last days of the USSR (E)
60. Assessing learning achievement (E)
61. Trends and developments in educational psychology (E)
62. A new partnership: indigenous peoples and the United Nations system (E/F/S)
63. Interpreting international comparisons of student achievement (E/F)
64. Education in the least developed countries: advancing in adversity (E/F)
65. Gender differences in learning achievement: evidence from cross-national surveys (E)
66. Développement de la culture scientifique et technologique dans l'éducation non formelle (F)
67. The pursuit of literacy (E)
68. Reforming schooling - what have we learnt? (E)
69. The role of measurement and evaluation in education policy (E)

* E = English, F = French, S = Spanish, A = Arabic, R = Russian

Educational studies and documents, 69





*The role
of measurement and evaluation
in education policy*

Edited by Frances M. Ottobre



The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Published in 1999 by the United Nations Educational,
Scientific and Cultural Organization
7, place de Fontenoy
75352 Paris 07 SP
Printed by UNESCO

ISBN 92-3-103605-1

© UNESCO 1999
Printed in France




Preface

This publication contains the papers and commentaries thereon presented at a round table of experts on the theme, 'The role of measurement and evaluation in education policy', that was organized by the International Association for Educational Assessment (IAEA), with financial support from UNESCO, at the headquarters of Educational Testing Service (ETS) in Princeton, New Jersey, United States of America, 29–30 June 1998. UNESCO's support for the round table was provided in the context of the preparation of the fifth edition of the Organization's *World Education Report*,

scheduled for publication in the year 2000.

UNESCO is grateful to both IAEA and ETS, as well as the participants in the round table, for their co-operation. In publishing the round table's papers and commentaries, the Organization considers that they merit the attention of a wider audience interested in the policy aspects of recent developments in education measurement and evaluation.


It should be noted that the opinions expressed in the papers and commentaries are the authors' and do not necessarily represent those of UNESCO.





Contents

| | |
|---|-----|
| List of acronyms | 8 |
| Foreword, Frances M. Ottobre | 9 |
| <i>Chapter 1. Key issues</i> , Samuel J. Messick | 11 |
| <i>Chapter 2. Equity in education and assessment</i> , Caroline Gipps | 15 |
| Comments on 'Equity in education and assessment' | 28 |
| Edmund W. Gordon | 28 |
| Pedro Ravela | 31 |
| Anil Kanjee | 34 |
| <i>Chapter 3. Education standards: current directions and implications for assessment</i> , | |
| Howard T. Everson | 39 |
| Comments on 'Education standards' | 49 |
| Gordon Ambach | 49 |
| Anton Luijten | 53 |
| Kent McGuire | 54 |
| <i>Chapter 4. Performance assessment</i> , Barry McGaw | 57 |
| Comments on 'Performance assessment' | 66 |
| Samuel J. Messick | 66 |
| Robert Linn | 69 |
| Mark D. Reckase | 71 |
| <i>Chapter 5. Purposes and challenges of international comparative assessments</i> , | |
| Tjeerd Plomp | 75 |
| <i>Chapter 6. International assessments: the United States TIMSS experience</i> , | |
| Albert E. Beaton | 89 |
| Comments on 'International comparative assessments' and 'The TIMSS experience' | 98 |
| Jahja Umar | 98 |
| Giray Berberoglu | 100 |
| <i>Chapter 7. Overview and synthesis: the role of measurement and evaluation in education policy</i> , Edmund W. Gordon | 103 |



List of acronyms

| | |
|--------|--|
| AAAS | American Association for the Advancement of Science |
| AAU | American Association of Universities |
| ACER | Australian Council for Educational Research |
| ACT | American College Admissions Test |
| CES | Civics Education Study |
| ESEA | Elementary and Secondary Education Act |
| ETS | Educational Testing Service |
| EU | European Union |
| GCE | General Certificate of Education |
| GCSE | General Certificate of Secondary Education |
| IAEA | International Association for Educational Assessment |
| IASA | Improving America's Schools Act |
| LEAs | Local Education Authorities |
| NAEP | National Assessment of Educational Progress |
| NAS | National Academy of Sciences |
| NCTE | National Council of Teachers of English |
| NCTM | National Council of Teachers of Mathematics |
| NFA | National Forum on Assessment |
| NRCs | National Research Co-ordinators |
| OECD | Organisation for Economic Co-operation and Development |
| OFSTED | Office for Standards in Education |
| SATs | Scholastic Assessment Tests |
| SIMS | Second International Mathematics Study |
| SITES | Second Information Technology in Education Study |
| TCI | Test Coverage Index |





Foreword

Frances M. Ottobre

Since 1976, IAEA has co-operated with UNESCO in the field of education measurement and evaluation, a field that has drawn increasing attention from national policy-makers in the follow-up to the World Conference on Education for All (Jomtien, Thailand, 1990). The round table featured in this volume sought to bring together distinguished experts in this field in order to examine selected issues for education policy that have emerged as the result of recent developments in the theory and practice of education measurement and evaluation. The financial support of UNESCO for the round table is gratefully acknowledged.

IAEA is also grateful to ETS, and in particular its President, Nancy Cole, for agreeing to host the round table. The late Samuel J. Messick, Distinguished Research Scientist at ETS, played a key role in organizing the round table and chairing its proceedings. His introductory remarks initiating the discussion at the round table are presented in the volume's first paper. There follow five papers that address key aspects of the theme of the round table, as well the commentaries on these papers that were prepared by selected participants. An 'Overview and Synthesis' of the discussion at the round table, based in part on Dr Messick's notes, is presented by Professor Edmund Gordon at the end of the volume.

IAEA joins with UNESCO in expressing its appreciation and thanks to all the participants for their thoughtful and insightful contributions to the round table: Gordon Ambach (Council of Chief State School Officers, United States); Albert E. Beaton (Boston College, United States), International Study Director of the Third International Mathematics and Science Study (TIMSS); Giray Berberoglu (Middle East Technical University, Turkey); Nancy Cole (ETS President, United States); Howard Everson (The College Board, United States); Caroline Gipps (University of London, United Kingdom); Edmund Gordon (John M. Musser Professor of Psychology Emeritus, Yale University, United States); Vincent Greaney (The World Bank); Anil Kanjee (Human Sciences Research Council, South Africa); Robert Linn (University of Colorado, United States); Anton Luijten (CITO, The Netherlands); Barry McGaw (Australian Council for Educational Research, Australia); Kent McGuire (Assistant Secretary of Education, United States); Samuel J. Messick (ETS, United States); Frances Ottobre (IAEA); Tjeerd Plomp (IEA President); Pedro Ravela (Proyecto de Mejoramiento de la Educacion de la Primaria, Uruguay); Mark Reckase (American College Testing Programme, United States); Alexander San-

nikov (UNESCO); Jahja Umar (Ministry of Education and Culture, Indonesia); Hans Wagemaker (Executive Director, IEA); and


Irene Walter (Caribbean Examinations Council).





1. Key issues

Samuel J. Messick



I will begin the round table by indicating some key issues that I think will be important in the course of our deliberations; I anticipate that others will be articulated as we have the opportunity to discuss the papers that will be presented. It will be useful, I think, to focus on the current uses of assessment and see what key issues emerge.

■ ■ **Uses of assessment** ■

A major use of assessment is to make decisions about instruction as it relates to individual students. Much of that use of assessment is part of teacher practice instead of education policy. However, there are some instances when very important education policy uses of assessment relate to individual students. These include using tests to decide whether students are eligible for needed services provided by state and federal entitlement programmes, and using tests to allocate education benefits and for deciding what form those benefits should take in individual student programmes.

There are important equity issues involved here. If you are going to make decisions about individual students, those decisions must be based on reliable and valid assessment. If they are not, it is

difficult to justify using tests to make decisions about students.

Another use of assessment is to provide information about the functioning of the education system such as the information that is provided by the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS).

A third use is to hold schools and educators accountable for student performance and to provide a lever for changing classroom instruction.

The final use that I will mention here is to certify specified levels of student achievement or mastery, as in tests for promotion, graduation and professional licensure.

■ ■ **Standards and equity** ■

Current education reform strategy stresses content standards and performance standards, standards-based assessment and score reporting, and performance assessment for higher order skills.

Equity issues surface once again. One such issue is the distinction between high standards for all students and common standards for all students. Until

recently, most of the rhetoric in the United States was based on common standards for all students. That is beginning to change and there is a subtle shift from common standards to high standards for all students. There is a growing awareness that standards may not need to be the same for all students and that individual differences of students may have some role in the education system. There is also growing recognition that content standards can be used to indicate what students should know and be able to do. Content standards that are defined for achievement will be deemed more important than those not listed and that could lead to the subtle possibility that the standards-based movement might cause, ironically, a restriction in the range of developed talent.

Low-stakes versus high-stakes assessment and the issue of equity

Low-stakes assessments provide information about performance to educators and to policy-makers, with no rewards or sanctions attached to the findings. It is assumed that information will motivate improvement, but there are potential negative consequences associated with low-stakes assessment, including the fact that the test information alone provided by low-stakes assessment may not convey clear implications for action. In recognition of this problem, NAEP and TIMSS have tried to collect additional information to add to the test data. Another potential negative consequence is that students may not be motivated to perform well on tests that have no consequences for them, thereby jeopardizing the validity of the scores.

High-stakes assessments provide information about performance, along with rewards and sanctions as spurs to action and accountability. But high-stakes assessments also have potential negative consequences associated with them. These include: teaching to the test and thus corrupting the indicator; narrowing the content and skills that

are taught in the classroom; widening the gap in education opportunities; and centralizing education decision-making and, consequently, deprofessionalizing teachers. That is why we must continue to pay a great deal of attention to these high-stakes tests to ensure that they are doing what they should be doing and to try to minimize the negative consequences associated with them.

Attribution of cause in high stakes assessment

One of the big issues we need to consider is how high-stakes assessments have persistently focused accountability on teachers and students. We must be concerned with the attribution of cause. The test score provides information not only about the performance of students, but also about the process and the people involved in the education process and the development of the test. Cause must be determined before anyone can be held accountable. It is unfair to hold someone accountable for something over which he or she has no control or responsibility.

In testing for promotion or graduation, where diplomas or salary increases can be withheld on the basis of test scores, it is essential that we be concerned about attribution of cause. We must ask ourselves whether it is the student's fault, the system's fault or the teacher's fault if the student failed. When a complicated system functions poorly, the finger pointing is usually in the direction of the weakest part. When a student is held culpable for his or her own failure, it may not be because he or she is really culpable; rather, it may be that he or she is the weakest part of the system. For this reason, we must develop principles for the attribution of cause. This important equity issue is frequently ignored. If you add the fact that the tests are combined with the standards-based reform movement and that concern for those high standards applies to the promotion or graduation tests for all students, the high

standards will lead to an unprecedented failure rate. So as we move forward in professing some of these policy issues, we must be alert to any anticipated consequences and try to take these issues into account.

Equality and equity

Another equity issue is how to distinguish fair assessment from equity in education and, in particular, how to distinguish equality from equity. There is a tension between key outcomes in education and equality of treatment. Many educators argue that the focus of key outcomes is not to indicate equal treatments. If we really want equal outcomes, will we be able to achieve them by maintaining equal treatments? Many in education, such as Dr Edmund Gordon, liken education to medicine. As in medicine, where the focus should be not on giving patients equal treatment but on giving them appropriate treatment, the focus in education should not be on treating students equally (that is, identically), but rather on treating students appropriately and sufficiently, based on the needs of each individual student. When can equal treatments be justified in standards-based education and assessment? These issues require additional research.

■ ■ ■ ■ **Large-scale assessment as policy research**

Large-scale assessment policy research can provide more information about processes and programmes that might convey clearer implications and action to help schools and teachers know what needs to be done to improve performance. Policy research is inherently anticipatory or predictive, but it is also contingent on variable and uncontrolled conditions. Policy predictions are, therefore, largely context-dependent and should be monitored in case conditions change. In fact, implementing a policy is an indication that a condition is going to change. We need to be aware of changing

conditions and we need to think about possible unanticipated consequences of initiating such changes.

Several years ago I listed the characteristics I thought would be important for large-scale assessments if they were to affect policy. I still believe that the following are important. In order to be effective policy instruments, large-scale assessments require:

- comparability of measures across time periods and population groups;
- interpretability of measures in terms of integrative constructs with predictive power;
- generalizability of measures across diverse contexts and background factors; or else the inverse (context dependence);
- relevance of performance measures to manipulable programme and process variables; and
- amenability to policy influence timeliness in response to pressing policy issues.

Policy analysis of such large-scale assessments involves:

- a developmental orientation to time-ordered data;
- empirically grounded construct interpretations of measures and relationships;
- inquiry into the role of diverse contributing factors; and
- appraisal of alternative perspectives on both the policy questions and the findings.

These are the kinds of characteristics Dr Albert Beaton and I tried to flesh out for TIMSS. Although we did the best we could considering the state of the art, it is clear that we need to do more to make large-scale assessments more useful to policy research so they can be more useful to policy-makers.

■ ■ ■ ■ **Summary**

As we proceed with the round table, we need to keep the importance and consequences of measurement and assessment in

mind, from the uses of student assessments for decisions about students and as a lever to change classroom instruction to their function in providing information about education systems, for purposes of promotion, graduation and professional licensure.

We must also keep in mind that the measurement process is the underpinning of standards-based education. Without a measurement scale there are no

performance standards. So measurement is critical for standards and for recording progress, but what gives power to the standards-based movement is performance assessment, especially that for higher order skills.


Finally, the importance of equity issues cannot be over emphasized. They permeate all the topics we will discuss during the next two days.





2. Equity in education and assessment

Caroline Gipps



■ Introduction

■
■
■

This paper addresses the implications of patterns of achievement for equity in policy and practice, specifically in relation to assessment and pedagogy. I shall address this issue using data from England and Wales, while also drawing on literature from a range of countries including the contributions to an international colloquium on issues of gender and pedagogy which Patricia Murphy and I organized for UNESCO in 1995. Indeed many of my views on these issues derive from discussions and joint work with Patricia Murphy (Gipps and Murphy, 1994; Murphy and Gipps, 1996). Much of the research on differential performance in the United Kingdom over the last fifteen years has focused on gender; there is much less detailed research on ethnic and cultural group differences in performance. Therefore, much of what I shall say relates to gender.

■ Equity issues

■
■
■

Prior to the recent change of government in the United Kingdom, equity had not been an underlying theme in education in

England and Wales. Debate and policy-making, where it has featured at all, has referred to equal opportunities in education, with an occasional mention of compensatory education for disadvantaged groups. Early attempts to achieve equality of opportunity for girls and boys focused in the main on equality of resources and access to curriculum offerings; important though this is, we now see it as a limited approach given the very different out-of-school experiences of girls and boys. The fundamental problem is that this policy focus reflects a deficit model approach to inequality: girls are 'blamed' for behaving like girls and encouraged to behave more like boys. The model implies the possibility of overcoming disadvantage through the acquisition of what is lacking. This approach leaves the status quo essentially unchanged, since girls are unlikely to achieve parity through equality of resources and formal equality of access alone. As Yates puts it, 'where the criteria of success and the norms of teaching and curriculum are still defined in terms of the already dominant group, that group is always likely to remain one step ahead' (Yates, 1985, p. 212). Equal opportunities is a policy area which was hotly contested in the United Kingdom in the 1980s: seen by the extreme right as a revolutionary

device which would disturb the 'natural' social order and as an attempt to attack White British society; and by the extreme left as essentially conservative because the gross disparities in wealth, power, and status which characterize our society remain unchallenged. A second approach is one which looks for equality of outcome (as evidence of equal opportunities), and this underpins analyses and discussions of group performance at public examination level in the United Kingdom.

The attitude to equity in the United States is very different from that in the United Kingdom, for reasons of history and because of the population structure. 'The US has a long-term commitment to equity for its wholly immigrant population' (Baker and O'Neil, 1994, p. 12), which is evidenced in equal outcome terms. 'The term equity is used principally to describe fair educational access for all students; more recent judicial interpretations, however, have begun the redefinition of equity to move toward the attainment of reasonably equal group outcomes' (Baker and O'Neil, 1994, p. 11). 'The educational equity principle should result in students receiving comparable education yielding comparable

performances' (Baker and O'Neil, 1994, p. 12), that is, equality of outcome.

Will Hutton, economist and editor of the *Observer* newspaper, points out that New Labour in the United Kingdom no longer aims for equality of outcome, preferring instead to work for equality of opportunity (*Observer*, 16 March 97, Here's a primary objective for New Labour). He argues, however, that this is a misguided approach, particularly in relation to the performance of primary schools, since we should not be content to see a wide variation of primary school performance. He argues that we must look for equality of outcome as the foundation for giving equality of opportunity for all school leavers, since progress in the primary phase is vital to later learning. This point of view is, however, more to do with expectations and intervention than with assessment.

Apple's (1989) review of public policy in the United States of America, Britain, and Australia led him to conclude that equality has been redefined: it is no longer linked to group oppression and disadvantage but is concerned with ensuring individual choice within a 'free market' perception of the educational community. In

Table 1. Curriculum and assessment questions in relation to equity

| Curricular questions | Assessment questions |
|---|---|
| Whose knowledge is taught? | What knowledge is assessed and equated with achievement? |
| Why is it taught in a particular way to this particular group? | Are the form, content and mode of assessment appropriate for different groups and individuals? |
| How do we enable the histories and cultures of people of colour, and of women, to be taught in responsible and responsive ways? | Is this range of cultural knowledge reflected in definitions of achievement? How does cultural knowledge mediate individuals' responses to assessment in ways which alter the construct being assessed? |

Source: C. Gipps and P. Murphy, *A Fair Test? Assessment, Achievement and Equity*, Milton Keynes, Open University Press, 1994; and after W. Apple, How Equality has been Redefined in the Conservative Restoration. In: W. Secada (ed.), *Equity and Education*, New York, Falmer Press, 1989.

Apple's view this redefinition has reinstated the disadvantage model and underachievement is once again the responsibility of the individual rather than the educational institution or community. He argues that attention in the equity and education debate must be refocused on important curricular questions, to which we add assessment questions (Table 1).

The view that Patricia Murphy and I take is that while one must strive to achieve actual equality of opportunity, equality of outcomes is not necessarily an appropriate goal: different groups may indeed have different qualities and abilities, and certainly experiences. Furthermore, manipulating test items and procedures in order to produce equal outcomes may be doing violence to the construct or skill being assessed and may camouflage genuine group differences (Gipps and Murphy, 1994). The concept of equity in assessment, as we use it, implies rather that assessment practice and interpretation of results are fair and just for all groups.

Willingham and Cole (1997), in a recent major report on gender issues and fairness in assessment, define test fairness as comparable assessment for each examinee. 'Fair test design implies comparable opportunity to demonstrate relevant knowledge and skills' (here, choice of construct and format are involved). 'In test development and administration, a fair test should provide comparable tasks, testing conditions, and scaled scores for all examinees' (here, selection of items/content, timing, computer use and any other conditions that might lead to anxiety are relevant). 'Fair test use should result in comparable treatment of examinees' (Willingham and Cole, 1997, p. 350).

Equity and assessment

It is important to remember that 'objective' assessment has traditionally been seen as an instrument of equity: the notion of the standardized test as a way of offering impartial

assessment is of course a powerful one, though if equality of educational opportunity does not precede the test, then the 'fairness' of this approach is called into question. Most attainment tests and examinations are amenable to coaching, and pupils who have very different school experiences are not equally prepared to compete in the same test situation.

As Madaus (1992) points out:

... in addressing the equity of alternative assessments in a high-stakes policy-driven exam system, policy must be crafted that creates first and foremost a level playing field for students and schools. Only then can the claim be made that a national examination system is an equitable technology for making decisions about individuals, schools or districts. (p. 32)

The same point is also made by Baker and O'Neil (1994).

Bias is a term widely used in relation to assessment and is generally taken to mean that the assessment is unfair to one particular group or another. This rather simple definition, however, belies the complexity of the underlying situation. Differential performance on a test (i.e. where different groups get different score levels) may not be the result of bias in the assessment; it may be due to real differences in performance among groups, which may in turn be due to differing access to learning, or to real differences in the group's attainment in the topic under consideration. The question of whether a test is biased or whether the group in question has a different underlying level of attainment is actually extremely difficult to answer. Wood (1987) describes these different factors as the opportunity to acquire talent (access issues) and the opportunity to show talent to good effect (fairness in the assessment). As Willingham and Cole (1997, p. 359) put it: 'Comparable opportunity to demonstrate skills is not the same as comparable opportunity to acquire skills'.

The traditional psychometric approach to testing operates on the assumption that technical solutions can be found to solve problems of equity with the empha-

sis on using 'elaborate' techniques to eliminate biased items (Goldstein, 1993; Murphy, 1990). A limitation of this approach is that it does not look at the way in which the subject matter is defined (i.e. the overall domain from which test items are chosen); nor at the initial choice of items from the thus-defined pool. Neither does it question what counts as achievement. It simply 'tinkers' with an established selection of items. Focusing on bias in tests, and statistical techniques for eliminating 'biased' items, not only may confound the construct being assessed, but has distracted attention from wider equity issues such as actual equality of access to learning, 'biased' curriculum, and inhibiting classroom practices.

When the existence of group differences in average performance on tests is taken to mean that the tests are biased, the assumption is that one group is not inherently less able than the other. However, the two groups may well have been subject to different environmental experiences or unequal access to the curriculum. This difference will be reflected in average test scores, but a test that reflects such unequal opportunity in its scores is not strictly speaking biased, though its use could be invalid. A considerable amount of effort over the years has gone into exploring cognitive deficits in girls in order to explain their poor performance on science tests. However, it was not until relatively recently that the question was asked whether the reliance on tasks and apparatus associated with middle-class White males could possibly have something to do with it. As Goldstein (1996) points out, tests are framed by the test developers' construct of the subject and their expectations of differential performance.

Fairness

Of course pupils do not come to school with identical experiences, and they do not have identical experiences at school. We cannot, therefore, expect assessment to have the

same meaning for all pupils. However, the stakes and purpose of the assessment are relevant here as Linn, Baker and Dunbar (1991) argue: 'On a non-threatening assessment such as . . . NAEP, for example, it is reasonable to include calculator-active problems even though student access to calculators may be quite inequitable. On the other hand, equitable access would be an important consideration in a calculator-active assessment used to hold students or teachers accountable' (p. 17). As Linn (1993) points out, the fairness of an assessment is an essential aspect of an overall judgement of validity. What is important is to have an equitable approach where the concerns, contexts and approaches of one group do not dominate. This, however, is by no means a simple task; for example, test developers may be told that they should avoid any context which may be more familiar to males than to females or to the dominant culture. But there are problems inherent in trying to remove context effects by doing away with passages that advantage males or females, because it reduces the amount of assessment material available. De-contextualized assessment is anyway not possible, and complex higher order skills require drawing on complex domain knowledge.

An alternative approach is to use, within any assessment programme, a range of assessment tasks involving a variety of contexts, a range of modes within the assessment, and a range of response formats and styles. This broadening of approach, though not always possible, is most likely to offer pupils alternative opportunities to demonstrate achievement if they are disadvantaged by any one particular assessment in the programme.

Indeed, this is included in the *Criteria for Evaluation of Student Assessment Systems* recommended by the National Forum on Assessment (NFA) (NFA, 1992, p. 32):

- To ensure fairness, students should have multiple opportunities to meet standards and should be able to meet them in different ways.

- Assessment information should be accompanied by information about access to the curriculum and about opportunities to meet the Standards.
- Assessment results should be one part of a system of multiple indicators of the quality of education.

Patterns of achievement

In England and Wales there are some clear differences in the performance of boys and girls, and in the performance of test-takers belonging to different ethnic groups.

Gender differences

In the National Curriculum Assessment, girls are outperforming boys in English and maths at ages 7, 11, and 14 (EOC/OFSTED, 1996). In the General Certificate of Secondary Education (GCSE), the exam taken at 16, girls gain more higher-grade (A, B, C) passes than do boys, a trend which has been growing since 1988. Since this trend coincided with the introduction of the GCSE, girls' superior performance was felt to be partly due to the style and approach of the new examination, which includes written examination papers and a performance assessment task carried out in school 'coursework'. Research shows, however, that course work contributes little to the final grade (Stobart et al., 1992), and that the exam papers and coursework have differential validity (Elwood, 1998). The GCSE exam did broaden the definition of achievement and the means of assessing it, while the introduction of the national curriculum meant that both genders had to study all subjects from ages 14 to 16, and it is likely that these two factors together have contributed to girls' growing success (Elwood, 1995). A similar overall pattern of improving female performance has also been found in the United States (Willingham and Cole, 1997), as well as links between gendered interests and performance.

Things have come a long way

since the days when young women were first admitted to examinations carried out by the London University Board. At first, the board insisted that young women were chaperoned and, in case the length of the examination proved too much for them, drinks were served and buckets of cold water were available in case any of them fainted (Kingdon quoted in Stobart et al., 1992).

At age 18+, in the pre-university Advanced or A-level exam, boys earn more higher (A, B, C) grades than girls earn, even in subjects in which girls did particularly well at age 16. For example, in English GCSE, girls have 13 per cent more higher-grades than do boys, while at 18 boys have 3 per cent more higher-grades than girls have in English Literature A level, despite making up only 30 per cent of the entry (Elwood and Comber, 1996). Longitudinal analysis of pupil performance from age 16 to age 18, using a new multi-level modelling technique, shows that for boys and girls with the same GCSE score, girls make less progress to A level, gaining around two points fewer than equivalent boys (Goldstein et al., 1997).

This pattern of performance prior to age 18 has contributed to anxiety about the performance of boys in the United Kingdom. A key factor is felt to be boys' lower motivation and more negative attitudes to school, particularly for working-class boys who can see reduced work opportunities for themselves in the changing labour market. Changing trends include growing part-time work, which women traditionally tend to take, and a growing service sector which requires high levels of communication skills at which women tend to be better. However, two points need to be made here: first, that the boys who go on to the General Certificate of Education (GCE) A-level do better than girls; and second, that gender, ethnic group, and social class interact to affect performance.

The research carried out by Patricia Murphy (1995) shows that girls and boys attend to different things in a task, and in this case neither response is wrong; both

responses are valid, just different. But there is evidence that one reason for boys' poorer performance is that tests and exams contain a greater verbal element than in the past, even in maths and science, and this is an area in which boys have always underperformed compared with girls. There is also evidence that boys tend not to use the sort of approaches to learning which current theories of learning advocate: relating knowledge to context in order to be able to apply it more widely; engaging in dialogue with other learners and the teacher in order to question and validate understanding; and using collaborative approaches to learning. These are some of the effective learning strategies which are more favoured by girls, and this tendency for girls to favour such strategies may go some way to explaining recent patterns of lower achievement by boys. Therefore, boys' approaches to learning may need to be reconsidered and reconstructed (Murphy and Gipps, 1996).

Ethnic group differences

A review of recent research on the achievement of ethnic minority pupils, commissioned by the Office for Standards in Education (OFSTED) (Gillborn and Gipps, 1996), shows that since the previous major review of ethnic minority performance more than 10 years previously (Swann, 1985), there were:

- generally higher levels of achievement, increasing year on year;
- improving levels of attainment among ethnic minority groups in many areas of the country; and
- dramatic increases in the examination performance of certain minority groups, even in school districts and local education authorities (LEAs) where there is significant poverty.

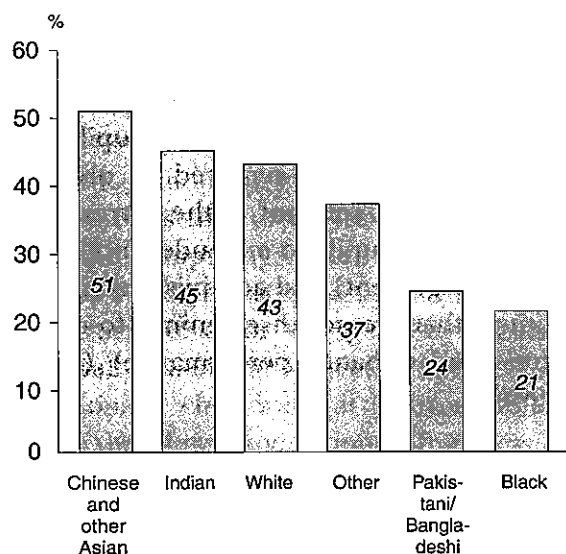
However, the gap is growing between the highest and lowest achieving ethnic groups in many LEAs. For example, African Caribbean young people, especially boys, have the lowest levels of performance; in

some areas their performance has actually worsened. The sharp rise in the number of exclusions from school also affects a disproportionately large number of Black pupils.

Research on the performance of infant and junior school (i.e. ages 6–11) pupils does not paint a clear picture: on average, African Caribbean pupils appear to achieve less well than Whites, although the situation is reversed in recent data from Birmingham. A more consistent pattern concerns the lower average attainments of Bangladeshi and Pakistani pupils in the early key stages: this may reflect the significance of levels of fluency in English, which are strongly associated with performance at this stage.

There are no up-to-date nationally representative figures on performance for 16-year-olds by different ethnic groups. However, the review of research and new LEA data identified some common patterns. Indian pupils also appear consistently to achieve more highly, on average, than pupils from other South Asian backgrounds. Indian pupils achieve higher average rates of success than their White counterparts in some (but not all) urban areas. There is no single pattern of achievement for Pakistani pupils, although they achieve less well than Whites in many areas. Bangladeshi pupils are known on average to have less fluency in English, and to experience greater levels of poverty, than other South Asian groups, and their relative achievements are often less than those of other ethnic groups. In one London borough, however, where resources and programmes have been focused on them, dramatic improvements in performance have been made and Bangladeshis are now the highest achieving of all major ethnic groups. African Caribbean pupils have not shared equally in the increasing rates of educational achievement: in many LEAs their average achievements are significantly lower than other groups. The achievements of African Caribbean young men are a particular cause for concern.

Figure 1: 15-16 year olds gaining 5 or more higher-grade GCSE passes by ethnic origin (England and Wales, 1994)



Source: 'Youth Cohort Study of England and Wales'. In: Office for National Statistics, *Social Focus on Ethnic Minorities*, London, HMSO, 1996.

Figures 1 and 2 show how complex the patterns of performance are. Both are taken from the Youth Cohort Study, (HMSO, 1996) which is the only nationally representative dataset to include information about class, gender and ethnic origin in England and Wales.

Figure 1 shows the proportion of different ethnic groups gaining 5 or more top grades (A, B, C) at GCSE at age 16, in 1994.

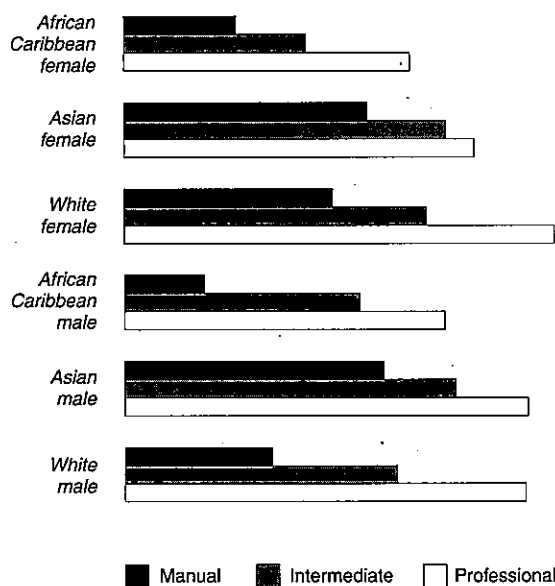
Figure 2 shows average exam scores by ethnic group, gender and social class in 1985 (more recent data containing all three factors are not available). This figure shows the complex interaction of social class, gender and ethnic origin: for example, only in the White group do girls consistently outscore boys, while among the Asian group that situation is reversed (although that pattern may have altered since 1985). Overall, the highest performing group is White girls from professional backgrounds and the lowest is African Caribbean males from a manual background.

The message of the review was that, where differences in performance are ignored and not monitored, patterns of inequality will increase. Looking on the constructive side, work carried out at LEA level indicates that where low performing ethnic minority groups are targeted sensitively with additional educational resources, those groups do perform at significantly higher levels (Gillborn and Gipps, 1996).

An agenda for assessment

So to return to our definition of equity, how do we ensure that assessment practice and interpretation of results are as fair as possible for all groups? It is likely that a wide-ranging review of curriculum and syllabus content, teacher attitudes to boys and girls and minority ethnic groups, assessment

Figure 2: Average exam scores by ethnic origin, gender and social class (England and Wales, 1985)



Source: D. Gilborn and C. Gipps, *Recent Research on the Achievements of Ethnic Minority Pupils*, London, OFSTED, 1996, adapted from D. Drew and J. Gray, *The Fifth Year Examination Achievements of Black Young People in England and Wales*, *Educational Research*, Vol. 32, No. 3, 1990, pp. 107-17, p. 114.

mode, and item format is required, as Table 1 shows, if we wish to make assessment as fair as possible. Although this is a major task, it is one which needs to be addressed given the growing importance of educational standards in many countries.

As Gipps and Murphy (1994) and Willingham and Cole (1997) argue, the construct tested is crucial. We need to encourage clearer articulation of the test/exam developers' constructs on which the assessment is based, so that the construct validity may be examined by test-takers and users. Test developers need to give a justification for inclusion of context and types of response mode in relation to the evidence we have about how this interacts with group differences and curriculum experience. The requirement is to select assessment content that accurately reflects the construct, even if it produces gender/ethnic group differences, and to avoid content that is not relevant to the construct and could affect such differences. The ethics of assessment demand that the constructs and the assessment criteria are made available to pupils and teachers, and that a range of tasks and assessments is included in an assessment programme. These requirements are consonant with enhancing construct validity in any case. Given the detailed and, as yet, poorly understood effect of context on performance, the evidence that girls more than boys attend to context in an assessment task, and the ability of changes in the context of the task to alter the construct being assessed, the area of construct validity demands detailed study. We certainly need to define the context of an assessment task and the underlying constructs, and make sure that they reflect what is taught.

We also need to encourage the continued use of a range of assessment modes and task styles. We also need to expand the range of indicators used: multiple indicators are essential so that those who are disadvantaged on one assessment have an opportunity to offer alternative evidence of their expertise (Linn, 1992, p. 44).

Caroline Gipps

If we wish pupils to do well in tests/exams we need to think about assessment which elicits an individual's best performance (Nuttall, 1987). Such assessments involve tasks that are concrete and within the experience of the pupil (an equal access issue), presented clearly (the pupil must understand what is required in order to perform well), relevant to the current concerns of the pupil (to engender motivation and engagement) and administered in conditions that are not threatening (to reduce stress and enhance performance) (Gipps, 1994).

Pedagogy

I now want to look at some of the issues around teaching and learning in school. Science and technology in Western cultures are seen as masculine subjects and they emerge as some of the most strongly sex-stereotyped areas of the curriculum. The overwhelming number of scientists who are male; the illustrations and examples used in teaching, through the world views, experiences, and ways of working that are assumed; and the way in which students and teachers in the laboratory context reconstitute gender in their interactions constitute a pedagogy for boys (Harding, 1996).

Classroom studies in England, the United States, and Australia indicate that in the area of computing:

- Boys are more confident than girls in their approach to work.
- The participation level in computer-use by girls (and female teachers) is lower than that of boys.
- Boys are also more likely than girls to use computers out of school, which contributes to their greater ease with them in school.

The difference in boys' and girls' confidence in the use of computers is also found in science: in both science and computer lessons boys are found to dominate the equipment and its use. In these subjects there is international evidence that boys

continue to have more positive attitudes than girls, although there are national exceptions. These differences can be linked to boys' greater experience with computers outside of school and with science-related activities (Littleton, 1996).

Research on mathematics performance among girls indicates that the picture mirrors that of the sciences, technology, and computing: where mathematics is seen as a preserve of males, girls choose not to participate in or engage with it, either physically when the subject is optional, or emotionally when it is not (Burton, 1996). Girls' own low expectations of success in mathematics, which are the product of cultural, family and societal pressures, can be self-fulfilling. Furthermore, the type of approach required to do advanced-level mathematical problem-solving requires girls to behave outside their typically socialized ways of behaviour: being independent, active, questioning and rule-breaking. Rather than blaming girls for not breaking the rules, we need to recognize that it is particularly difficult to break rules when one is socialized to be compliant (Walden and Walkerdine, 1985). But it is not only girls who may find school mathematics inaccessible:

the mathematics curriculum has tended to emphasize values and concerns which are more middle class than working class, and to draw on experiences which are more relevant to children of Anglo-Celtic descent (Australian Education Council, 1991).

There is a weight of research evidence to indicate that girls do extremely well in reading and writing in their language of instruction. This simple statement, however, masks a complex of issues. First, boys' underachievement in the subject is excused by downgrading that subject (Cohen, 1996). Extraordinary though it may seem, the argument in England has been that boys' poor performance in English language studies at school does not matter, since English is a girls' subject anyway, and boys get on well without it. The latter point is, of course, true, and we must ask why this lack of suc-

cess seems to be of so little hindrance. Boys' competence in oral language has been thought to be one factor, although more recent evidence from public exam data in England suggests that girls are outperforming boys on the assessment of oracy. The nature of their out-of-school reading is another: comics and fact-based books which boys prefer prepare them for engaging with text books and support their development of a scientific style of writing (White, 1996)

Teachers' understandings, beliefs and expectations, together with those of students, are crucial in schooling. These define and normalize what is considered to be appropriate, reasonable and effective for different groups and categories of pupils. The role of language is a major theme in the construction of knowledge and meaning; the many pupils who do not speak the language of instruction, or indeed do not share the dialect and social and cultural mores of the dominant educator group, are therefore at a disadvantage.

Interventions

I now outline some of the findings of intervention studies that have focused on girls' performance (and we are now using similar strategies to enhance boys' performance). Some interventions focus on separating gender groups for longer or shorter periods in order to offer 'space' to the girls, tutoring on their own terms for both groups, an opportunity to reflect on the values and attitudes of the other sex, and on working together in mixed classrooms. It is the alternating between single-sex and co-ed settings, and reflecting on differences, which seem to be the powerful factors. In this way boys, as well as girls, are brought into discussion about knowledge and gendered-appropriate or inappropriate behaviour.

But we cannot separate the 'how' of learning from the 'what'. In Western cultures, a successful pedagogy for girls in science and technology, different from

the traditional approach appropriate mainly for males, has been identified. This places teaching/learning in a social context, relating to human needs and 'real' problems; and allows for collaborative ways of learning, including discussion-based exploration of understanding, as well as providing assessment procedures that allow for the recognition of complexity and the identification of a range of problems (Harding, 1996). A network of science educators in Victoria, Australia has challenged both the definition of the physics curriculum and its teaching and assessment practices. A programme in which physics is learnt in context and assessed innovatively has enhanced the performance of both boys *and* girls, but has in particular elicited excellence from girls (although not those from lower socio-economic status (SES) families, see Hildebrand (1996).

At tertiary level, too, successful interventions need to focus on a range of issues together. These include addressing the knowledge content of the curriculum and the way in which knowledge is contextualized and presented: a content-driven curriculum with a teaching-as-telling pedagogy is the norm in physical and engineering sciences. Presenting the curriculum in a 'gender-inclusive' way, that is in a more connected style with social and environmental contexts integrated, can attract a more diverse range of students to study science at tertiary level and will diversify the culture of science (Lewis, 1996).

The research reviews and intervention studies suggest that to enhance the performance and engagement of lower performing groups – be it girls in science and mathematics, boys in language, or low-performing ethnic minorities – we need to examine the knowledge base of the curriculum being offered, as well as how that knowledge is taught.

We know from intervention studies that in order to enhance the performance of girls, teachers need to be made aware of and encouraged to use variations in teaching strategies by:

- using more co-operative and interactive modes of learning;
- emphasizing discussion and collaboration;
- having class discussion and quiet reflection;
- using 'private' as well as public questioning and probing of the pupil by the teacher;
- slowing the pace of a lesson and encouraging pupils to use the time to compose responses; and
- giving feedback which challenges and gives precise guidance (in a supportive manner) as well as praise, rather than the bland praise (for dutiful hard work), which girls currently tend to receive. (All learners need to be given encouragement to go beyond that which is known and to undertake the exploration of new ideas.)

Interventions in curriculum and pedagogy will not, however, alter the pattern of success unless the assessment system is also changed to be consistent with the heterogeneity of the learning population. Using a range of assessment processes, together with clarity and openness about what is being assessed and how, is not only more equitable, but also supports learning (Gipps and Murphy, 1994).

In good assessment practice we should:

- use assessment that supports learning and reflection, including formative assessment;
- design assessment that is open and linked to clear criteria rather than relying upon competition with others; and
- include a range of assessment strategies so that all learners have a chance to perform well.

We need to think of pedagogy as being composed of a range of strategies (which includes a variety of materials and content, teaching styles, and classroom arrangements/rules) for different groups of pupils and for different subject areas. In the famous words of Ausubel (1968) 'the most important single factor influencing learn-

ing is what the learner already knows. Ascertain this and teach him accordingly'. This instructs the teacher to focus on the learner's understanding, but we now know that it is not sufficiently encompassing. What the learner knows is itself a function of context, learning style, materials and classroom interaction, all of which are deeply affected by gender, ethnic group and social class.

Pupils too need to understand that there is a range of learning strategies which are appropriate for different tasks, subjects and purposes. They also must learn to choose the appropriate learning strategy for a particular setting/occasion. This resonates with what we know about metacognition: that pupils need to be aware of and to monitor, 'to regulate', their own process of learning. The emphasis here, just as with the teacher, is on the pupil as a conscious decision-maker.

Most of this section on pedagogy is related to gender, with an occasional reference to social class/disadvantage. Little work has been done, at least in the United Kingdom, on differences in learning and reactions to curriculum texts among various ethnic groups. The low performance of African Caribbean boys in England, in particular, suggests that such work is now overdue. It has been suggested that in the United Kingdom many teachers have negative attitudes towards African Caribbean pupils and low expectations of their academic performance; that as a result, they treat them less favourably in the classroom and in wider school processes, denying them the educational opportunities enjoyed by their White peers; and that many, therefore, experience alienation, low academic achievement and, consequently, restricted life chances (Gillborn, 1995). This is less likely to be the case for Asian pupils, whose families are perceived as valuing education, whose approach to schooling is more similar to that of the White middle class, and for whom any early problems may be seen to be related to lack of proficiency in English.

Conclusions

Early work on the education of girls concentrated on 'changing girls' (to persuade them to engage with science, etc.). The approach then switched to 'changing subjects' (challenging the traditional curriculum and views of knowledge). We are now moving into a phase of changing and diversifying our pedagogic and assessment strategies to suit a range of learners, in order to cater for a range of ways of learning and knowing.

This third approach results from a range of shifts in thinking:

- from a post-modernist critique of one overarching feminist approach which denies differences among girls;
- from a post-structuralist challenging of the 'objective' reality of science and maths;
- from understandings that clever/strong girls who achieve in maths/science make a positive choice not to pursue these beyond school (or university);
- from understandings that the curriculum subject matter, material and teaching approaches have not engaged boys or girls who are not White and middle class; and
- from developments in cognition and learning theory that tell us to respect and engage with the learner.

Pupils do not all learn in the same way and a class of pupils will need different strategies. This is similar to the argument we make in relation to assessment: if genuine equality of access is a prime requirement for equity, then in any assessment programme we need to include a range of content, context and types of task and response modes so as to offer all groups an opportunity to perform well.

Changing teachers' approaches so that they consider a range of pedagogic strategies appropriately for various pupils, subjects and tasks places a tremendous demand on teachers and on how they are educated. But the task is not just about

gender; it is a much broader agenda of engaging with the learner, while being conscious of the 'White, male, middle-class' nature of knowledge as it is defined, so as to offer appropriate and effective teaching for *all* groups of pupils. In Western countries ethnic minority and disadvantaged boys *and* girls are underachieving. While there is evidence that strategies to make the curriculum and teaching more 'girl friendly' have worked with girls from majority, middle-class backgrounds, they have not worked with other girls; furthermore, they have often generated a 'male backlash.' As Kenway, Blackmore, Willis and Rennie (1996) argue, we are in an age of complex, shifting social and cultural circumstances with many males (and not just those for whom manual jobs were/would be the normal expectation) feeling threatened. Alternative ways of expressing their masculinity/power include violence and scapegoating, hence the growing harassment of girls and women.

Although we do not necessarily expect equality of outcome, we must continue to seek genuine equality of access. This means that all courses, subjects studied, examinations and so forth, should be equally available to all groups and should be presented in such a way that all groups feel able to participate fully. One suggestion from the United States is that, since opportunity to learn is a key factor in performance, schools may have to 'certify delivery standards' as part of a system for monitoring instructional experiences (Linn, 1993). How realistic it is to do this remains to be seen, but it does put the onus on schools to address the issue of equal access, at an actual rather than formal level.

However, no discussion of assessment, equity and pedagogy can be framed outside the political context: we must look at teaching and learning in relation to the education-gender-ethnic system. The micro-politics of the class and the school are also very powerful in the education experiences of different groups of pupils. We have to accept that research is

limited in the impact it can make, without the political will to support changes to the system, and in particular, to the status quo.

References

- APPLE, M. W. 1989. How Equality has been Redefined in the Conservative Restoration. In: W. Secada (ed.), *Equity and Education*. New York, Falmer Press.
- AUSUBEL, D. P. 1968. *Educational Psychology: A Cognitive View*. New York, Holt, Reinhardt and Winston.
- AUSTRALIAN EDUCATION COUNCIL. 1991. *A National Statement on Mathematics for Australian Schools*. Victoria, Curriculum Corporation.
- BAKER, E.; O'NEIL, H. 1994. Performance Assessment and Equity: A View from the United States of America. *Assessment in Education*, Vol. 1, pp. 11-26.
- BURTON, L. 1996. A Socially Just Pedagogy for the Teaching of Mathematics. In: P. Murphy and C. Gipps (eds.), *Equity in the Classroom: Towards Effective Pedagogy for Girls and Boys*, pp. 136-45, London/Paris, Falmer Press/UNESCO Publishing.
- COHEN, M. 1996. Is There a Space for the Achieving Girl? In: P. Murphy and C. Gipps (eds.), *Equity in the Classroom: Towards Effective Pedagogy for Girls and Boys*, pp. 124-35. London/Paris, Falmer Press/UNESCO Publishing.
- DREW, D.; GRAY, J. 1990. The Fifth Year Examination Achievements of Black Young People in England and Wales. *Educational Research*, Vol. 32, No. 3, pp. 107-17.
- ELWOOD, J. 1995. Undermining Gender Stereotypes: Examination and Coursework Performance in the United Kingdom at 16. *Assessment in Education*, Vol. 2, No. 3, pp. 283-303.
- . 1998. Testing and Evaluation: Confronting the Challenges of Rapid Social Change. Conference on equity issues in performance assessment: The contribution of teacher-assessed

coursework to gender-related differences in examination performance, Barbados.

- ELWOOD, J.; COMBER, C. 1996. *Gender Differences in Examinations at 18+*. London, Nuffield Foundation/University of London Institute of Education.
- EOC/OFSTED. 1996. *The Gender Divide: Performance Differences between Boys and Girls at Schools*. London, Office of New Majesty's Chief Inspector of Schools and the Equal Opportunities Commission.
- GILLBORN, D. 1995. *Racism and Antiracism in Real Schools*. Buckingham, Open University Press.
- GILLBORN, D.; GIPPS, C. 1996. *Recent Research on the Achievements of Ethnic Minority Pupils*. London, OFSTED.
- GIPPS, C. 1994. Developments in Educational Assessment: What Makes a Good Test? *Assessment in Education*, Vol. 1, No. 3, pp. 283-91.
- GIPPS, C.; MURPHY, P. 1994. *A Fair Test? Assessment, Achievement and Equity*. Milton Keynes, Open University Press.
- GOLDSTEIN, H. 1993. Assessing Group Differences. *Oxford Review of Education*, Vol. XIX, pp. 141-50.
- . 1996. Group Differences and Bias in Assessment. In: H. Goldstein and T. Louis (eds.), *Assessment in Society: Problems, Developments and Statistical Issues*. Bognor Regis (United Kingdom), Wiley. (Probability and Mathematical Statistics.)
- GOLDSTEIN, H.; THOMAS, S.; O'DONOGHUE, C.; KNIGHT, T. 1997. *DFEE Study of Value Added for 16-18 year olds in England*. London, DFEE.
- HARDING, J. 1996. Girls' Achievement in Science and Technology: Implications for Pedagogy. In: P. Murphy and C. Gipps (eds.), *Equity in the Classroom: Towards Effective Pedagogy for Girls and Boys*, pp. 111-23. London/Paris, Falmer Press/UNESCO Publishing.
- HILDEBRAND, G. 1996. Redefining Achievement. In: P. Murphy and C. Gipps (eds.), *Equity in the Classroom: Towards Effective Pedagogy for Girls and Boys*, pp. 149-72. London/Paris, Falmer Press/UNESCO Publishing.
- KENWAY, J.; BLACKMORE, J.; WILLIS, S.; RENNIE, L. 1996. The Emotional Dimensions of Feminist Pedagogy in Schools. In: P. Murphy and C. Gipps (eds.), *Equity in the Classroom: Towards Effective Pedagogy for Girls and Boys*, pp. 242-60. London/Paris, Falmer Press/UNESCO Publishing.
- LEWIS, S. 1996. Intervention Programs in Science and Engineering Education: From Secondary Schools to Universities. In: P. Murphy and C. Gipps (eds.), *Equity in the Classroom: Towards Effective Pedagogy for Girls and Boys*, pp. 192-213. London/Paris, Falmer Press/UNESCO Publishing.
- LINN, M. C. 1992. *Gender Differences in Educational Achievement. Sex Equity in Educational Opportunity, Achievement and Testing*. Princeton, Educational Testing Service.
- LINN, R. L. 1993. Educational Assessment: Expanded Expectations and Challenges. *Educational Evaluation and Policy Analysis*, Vol. XV, Nos. 1-16. (ERIC No. EJ463413.)
- LINN, R. L.; BAKER, E.; DUNBAR, S. 1991. Complex, Performance-Based Assessment: Expectations and Validation Criteria. *Educational Researcher*, Vol XX, No. 8, pp. 15-21.
- LITTLETON, K. 1996. Girls and Information Technology. In: P. Murphy and C. Gipps (eds.), *Equity in the Classroom: Towards Effective Pedagogy for Girls and Boys*, pp. 81-96. London/Paris, Falmer Press/UNESCO Publishing.
- MADAUS, G. 1992. *A Technological and Historical Consideration of Equity Issues Associated with Proposals to Change the Nation's Testing Policy*. Symposium on Equity and Educational Testing and Assessment, Washington, DC.
- MURPHY, P. 1990. *Gender Differences: Implications for Assessment and Curriculum Planning*. Roehampton, British Educational Research Association (BERA).

- . 1995. Assessment-gender Implications. In: D. Farrelly (ed.), *Examinations in the Context of Change*. Dublin, University College.
- MURPHY, P.; GIPPS, C. (eds.). 1996. *Equity in the Classroom: Towards Effective Pedagogy for Girls and Boys*. London/Paris, Falmer Press/UNESCO Publishing.
- NATIONAL FORUM ON ASSESSMENT. 1992. Criteria for Evaluation of Student Assessment Systems. *Educational Measurement: Issues and Practice*, Spring.
- NUTTALL, D. 1987. The Validity of Assessments. *European Journal of Psychology of Education*, Vol. XI, No. 2, pp. 108–18.
- SMITH, P.; WHETTON, C. 1988. Bias Reduction in Test Development. *The Psychologist*, July, pp. 257–58.
- STOBART, G.; ELWOOD, J.; QUINLAN, M. 1992. Gender Bias in Examinations: How Equal are the Opportunities? *British Educational Research Journal*, Vol. XVIII, No. 3, pp. 261–76.
- SWANN, M. 1985. *Education for All*. London, Her Majesty's Stationery Office.
- WALDEN, R.; WALKERDINE, V. 1985. *Girls and Mathematics: From Primary to Secondary Schooling*. London, Institute of Education. (Bedford Way paper. Rep. 24.)
- WHITE, J. 1996. Research on English and the Teaching of Girls. In: P. Murphy and C. Gipps (eds.), *Equity in the Classroom: Towards Effective Pedagogy for Girls and Boys*, pp. 97–111. London/Paris, Falmer Press/UNESCO Publishing.
- WILLINGHAM, W.; COLE, N. 1997. *Gender and Fair Assessment*. London, Lawrence Erlbaum Associates.
- WOOD, R. 1987. *Measurement and Assessment in Education and Psychology*. London, Falmer Press.
- YATES, L. 1985. Is 'Girl Friendly Schooling' Really What Girls Need? In: J. Whyte, R. Deem, L. Kant and M. Cruickshank (eds.), *Girl Friendly Schooling*. London, Methuen.

Comments on 'Equity in education and assessment'

Edmund Gordon

28

Introduction

I agree with many of the values reflected in Gipps' paper, but I wish that she had elaborated more extensively on the complexities of the pedagogical and psychometric issues relevant to this topic. Her use of data from studies of differential performance related to gender focuses us in the right direction, that is, on the grounding of the problems of equity in education and assessment in issues related to status and human social divisions. One immediately recognizes that

class, ethnicity, dominant language proficiency and so on could be substituted for gender without greatly changing the basic thesis of the paper. Gipps reminds us that the issues are complicated, but the nature of this complexity is insufficiently elaborated. My remarks address some of the complexities of the relationship between membership in one or more of these social divisions and academic achievement.

We recognize the well-established association between low status and academic achievement. Whether the indicator is caste, class, gender, national origin

Edmund Gordon

or race, when the status assigned to the group is low, there is a strong tendency for members of the group to perform lower on tests of academic achievement, compared to how their more privileged peers perform on the same tests. Debate continues relative to whether this difference is a result of the status of such group members and the way they are treated by the social order, or if the culpable factors are to be found in the behaviours of the persons themselves. It is likely that both of these sources of influence are involved.

■ Causal relationships ■

Some clarity concerning possible causal relationships between membership in one of the social divisions and academic achievement may come from the distinctions that can be made between the status and the functional characteristics of group members. With respect to education, there appear to be few characteristics that are intrinsic to one's caste, class, gender or race status that can explain the persistent differences observed in school and academic test performance. One's status may be more associated with how one is treated, the nature of one's access to education and what is expected of the person or group than with what is possible. Thus the growing concern with the provision of equitable opportunities to learn. On the other hand, the functional characteristics – what one does or is enabled to do, the way in which one behaves or is expected to behave – are thought to have powerful effects on learning and ultimately on academic achievement. This conception of possible cause is reflected in current emphases on academic socialization and effort. Hypotheses implicit in these conceptions have not been subjected to rigorous tests.

These and other issues related to differentials in the academic achievement of different social groups suggest that problems of education assessment may be secondary to the problems of access to the

essential prior conditions of academic achievement (i.e. capital resources necessary for human development) and meaningful engagement in effective teaching and learning transactions. Coleman, Hoffer, Lockheed, Miller, Gordon and Meroe have concluded that it may well be that pedagogy is enabled to work, in part, because of the availability of such capital invested in the development of the learner and her or his education. Among the categories of human development capital that have been described are the following:

Health capital: physical developmental integrity, health and nutritional condition, etc.

Financial capital: income and wealth, family, community and societal economic resources available for education.

Human capital: social competence, tacit knowledge and other family related education advantages.

Social capital: social network relationships, social norms, cultural styles and values.

Polity capital: societal membership, social concern, public commitment, political/economic power.

Personal capital: disposition, attitudes, aspirations, efficacy, sense of power and the deployment of effort.

Institutional capital: quality of and access to educating and socializing institutions.

Pedagogical capital: supports for appropriate educational treatment in family, school and community.

Almost four decades ago John Carrol suggested that aptitude might best be considered a function of devoting sufficient time to learning tasks that are appropriate to what must be learned. I have long believed that if we are to take Carrol seriously, we will need to give more attention in teaching and learning to 'time spent on appropriate learning tasks', learner engagement and persistence with those tasks, learner resource utilization and the learners' sense of the efficacy of the learning effort. It can be argued that when these prior conditions are adequate the equity issues in teaching, learning and assessment

become manageable. It is then that we can productively attend to the fine tuning of education practice, assessment and policy that may be necessary to more effectively address the problems of equity in education and assessment. I assert that the problems of education assessment may be secondary to the problems of learner access to the resources necessary for optimal human development, and access to effective teaching and learning experiences. Issues of equity in assessment are less critical in the presence of universally effective education. However, this does not absolve the measurement community of responsibility for being responsive to issues of equity.

Four categories of equity issues

I have categorized equity issues in education assessment into four broad categories: (1) the political economy of education assessment; (2) limitations in the technical capacities of pedagogy and its assessment; (3) the epistemological and theoretical contexts of education measurement; and (4) the technological demands of equitable systems of education assessment.

With respect to the political economy of education assessment, access to a wide variety of human resource capital, including the opportunity to learn, is a condition that is essential to any consideration of equity and fairness in education assessment. It is foolhardy to assume that advances in the technology of measurement can make up for deprivation of resource capital and adequate opportunities to learn. Traditionally these contextual factors would be of no concern to education measurement; but education measurement should be concerned with the assessment of learner status and function, since it is the understanding of both learner status and functional processes that informs pedagogical decisions and intervention. Thus the equity issue that flows from the capital resource factors associated with the political econ-

omy of assessment has to do with the complementarity of topographical and topological analysis of learners' status and function.

Political and professional limitations in the capacity of the education and education assessment enterprises are identified as impediments to equitable assessment. These may be reflected in the mismatch between the diverse conditions and characteristics of learners, the pluralistic demands of modern societies and the professional competencies of educators and assessors, as well as in the absence of political and professional will to make available and apply the best of what we know to segments of the population where it is especially needed.

I could use all of our time together to discuss the 'match/mismatch problem,' but instead I will refer to one line of investigation. Thirty years ago, Irwin Katz, called attention to differences in responses to our tests; these differences were associated with ethnic differences in examiners and examinees, and among test-takers. Fifteen or twenty years ago Sam Messick and his colleagues were studying the negative impact of hostile environments on test-takers. Now, in the mid and late 1990s, Claude Steele is making headlines with his research indicating that minority students perform differently under different ethnicity-related conditions of performance demand. We have not learned how to deal with these conditional correlates of human performance. Rather, we continue to measure developed abilities as rarified and autonomous components, as if they were simply intrinsic to the persons being assessed. The equity issues have to do with how we factor in and accommodate conditional and situational correlates of human performance and how we generate the professional and public will to apply such knowledge to our work.

The epistemological and theoretical contexts of education assessment are identified as sources of concern reflecting the need for a greater confluence of knowledge and technique flowing from the sci-

ences of mental development and learning and the sciences of education measurement. If mental development and learning are best understood as involving affective, cognitive and situative processes, measurement can not be limited to on-demand recall of iconic representations of knowledge, techniques and values drawn from a truncated and hegemonic canon. If the development of intellectual competence is the primary function of pedagogy, and such competence is thought to be reflected in the capacity to adapt available resource information to solving routine and novel problems, then the assessment of education outcomes must provide learners with challenges and opportunities to demonstrate such competence. The equity issue here corresponds to one of the central issues in assessment: how do we achieve greater symmetry between changing conceptions of knowledge, changing conceptions of pedagogy, changing conceptions of assessment, and changing conceptions of intellectual competence? Our traditional technology of assessment is increasingly challenged by these shifting conceptions.

The fourth equity issue concerns the generation of a strategic plan for the development of equitable systems of education assessment in which equity and excellence are privileged. In an earlier work, I suggested several initiatives that could move the field toward a more equitable system of education assessment:

- diversity in teaching, learning and assess-

- ment experiences, including tasks, contents, contexts, demands and referents;
- flexibility in the timing of teaching, learning and assessment entry points, and in the time spans allowed for learning and performance;
- multiplicity in the perspectives to which students are exposed, as well as in the perspectives which students are encouraged to express, and that are accepted, with the requirement that the students engage in comparison and justification;
- critical sampling from canonical and non-canonical views, knowledge and techniques;
- the use of hypertext (i.e. imbedded substantive and/or procedural knowledge), with the requirement that the absent element be provided;
- choice involving self-selected and teacher/examiner-selected options for the demonstration of what is known;
- opportunity to identify in the indigenous experience examples of canonical knowledge and technique, and in the canonical find examples of the indigenous experience;
- individual and co-operative learning and performance opportunities; and
- self-designated tasks from learner/examinee-generated inventories of knowledge, skill and understanding: what do I know or want to know, and how do I choose to learn and demonstrate that I have learned it?

Comments on 'Equity in education and assessment'

Pedro Ravela

I found Gipps' paper interesting and enlightening. In my discussion, I will begin by focus-

ing on three main aspects of the paper. Then, I will share some different points of



view about equity in assessment and education that are related to South America and, particularly, to Uruguay.

Culture, assessment and pedagogy

First, Gipps makes the point that if we do take the cultural background of different groups into account, the differences measured by our tests may be the consequence of the type of tasks included in our instruments, rather than the consequence of different levels of mastery and attainment of students. I agree with the statement that activities proposed in tests should reflect different approaches to knowledge and should allow students with different cultural backgrounds to demonstrate their abilities. Test activities should not be more appropriate for some cultural backgrounds than for others.

But developing tests that meet these conditions is not an easy task. While it is possible to avoid significant bias in test tasks, it is quite difficult to build a test that is completely appropriate for everyone. If we were to take this goal to its extreme, we would have to design a different test for each student.

Second, I entirely agree with the paper when it emphasizes the need to clarify what is assessed. The construct and criteria for designing tasks must be explicit. I am convinced that there must be considerable discussion around the issue of what skills and knowledge should be assessed, and that the construct must reflect consensus about the fundamental goals of the education system. Further, assessment results must be examined with the understanding that their meaning depends on the social relevance of the abilities and knowledge that are evaluated.

Third, I appreciate the concern expressed for the necessity of different pedagogical approaches appropriate to different groups. The statements about differences between boys and girls in learning sci-

ences and language were enlightening. The paper illustrates the kind of education research we need to carry out regarding how different people learn and how they should be taught.

Equity in developing countries

Now, I would like to make some comments about equity, from a different point of view, regarding the reality of developing countries, especially in South America.

Developing countries have only recently begun to implement education assessment systems. Most of the assessment is on the education system. This means that test results have no consequences on, or will not affect students' careers and lives. This fact makes a difference in how educators in these developing countries think about equity. In this context, equity assumes a different meaning. The focus is not on fairness in examinations. Rather, our main concern about equity is how the assessment system can contribute to guaranteeing that the entire population achieves mastery of fundamental skills and knowledge. We define as fundamental those abilities that people need to continue to learn, to understand society, to participate as citizens and to have the opportunity to work in a satisfying job. Our first challenge is not about how to respect cultural differences among groups, but about how we ensure that everyone achieves mastery of these fundamental abilities. For this reason, I cannot agree with Gipps when she states:

While one must strive to achieve actual equality of opportunity, equality of outcomes is not necessarily an appropriate goal: different groups may indeed have different qualities and abilities, and certainly, experiences.

I think it is a bit risky to state that all differences are genuine differences, that we must respect them, that we must not attempt equality of outcomes and that we should

align the assessment system with the differences in the learning population.

In developing countries (and particularly in Uruguay, which doesn't have an indigenous population), cultural differences in society are mainly a result of poverty. Some 38 per cent of mothers in Uruguay have not had the opportunity to study beyond elementary school. The children of these women have great difficulty developing fundamental abilities like reading and writing. For example, in Uruguay's primary schools, the average-repetition rate for first-graders is 22 per cent. In Uruguay's poor neighbourhoods, this rate is more than 30 per cent. This means that one of out every three children fails. In this context, I cannot say that these children are just different because they don't use written language and, consequently, change my tests to an oral mode. I cannot accept that children from those families cannot learn to read and write. These differences are a result of lack of opportunity. Education assessment systems must highlight these types of differences, so that we can work to reduce the gap between social groups. There may never be total equality of outcomes. People develop different abilities. This is not a problem as long as we assure that everyone has first achieved fundamental skills.

Assessment's main concern must be the skills and knowledge that are truly fundamental for understanding and participating in society. An excessive concern about respect for differences may leave certain groups without the skills needed to participate in a complex world. Assessments must not only be fair to different cultures and groups, but also contribute to the aim that social groups with different backgrounds attain similar levels of achievement in the most relevant skills and knowledge. Consequently, I agree again with the paper when it states 'the requirement is to select assessment content that accurately reflects the construct, even if it produces gender/ethnic [I add social-group] differences, and to avoid content that is not relevant to

the construct and could affect such differences. The ethics of assessment demand that the construct and assessment criteria are made available to pupils and teachers.'

However, what follows is, 'We certainly need to define the context of an assessment task and the underlying constructs, and make sure they reflect what is taught.' Some of us have a different point of view. In Uruguay we use assessment to orient teachers about what to teach. Through the discussion about the construct to be assessed, and through the dissemination of assessment results, we give teachers some guidance regarding the fundamental skills and knowledge our students need to attain. In other words, we use the assessment to 'teach the teacher what to teach'.

It is important to reiterate that in Uruguay the assessment will not affect the student's career. Therefore, equity for us is not only a matter of guaranteed fairness among individuals, but also the main way to guarantee that everyone achieves the basic skills.

Finally, another sense of equity in assessment that I am convinced is very important is related to the reporting of results. Most South American countries report results in one dimension: average test scores of public and private schools, average test scores of different provinces or states, average test scores of single schools. The same is true with international studies, according to some of the reports I have seen. The social differences between public and private schools' populations, or the social composition of populations in different provinces or countries, does not enter into the reports.

A study in Uruguay in 1996 carried out a national assessment in language and math at the last grade of primary education, age 11. In private schools, 61 per cent of students achieved mastery in math, while in public schools only 35 per cent of students achieved mastery. The first, and wrong, conclusion is that private schools are better than public schools. The conclusion

is wrong, based on the information obtained from a social questionnaire sent to every student's home. In that questionnaire we asked questions regarding parent's education and occupation, number of people living at home, home equipment, number of books at home, and so on. Parents received the questionnaire the first day, and sent it back to school the next day, in a closed envelope. Some 98 per cent of the families responded.

On the basis of these data, we classified schools into five social context categories: 'very favourable context', 'favourable context', 'medium context', 'unfavourable context' and 'very unfavourable context'. When we compared private and public schools within each of the categories of social context, academic differences disappeared. For example, in 'very favourable

context' schools, 71 per cent of students attended private schools and 66 per cent of students attend public schools. The gap was reduced to 5 points. In 'favourable' and 'unfavourable' contexts, students attending public schools achieved better than those attending private schools – 46 per cent to 41 per cent in favourable contexts and 23 per cent to 15 per cent in unfavourable contexts. It is very likely that something similar happens with international studies. But, what kind of population is behind results? What is the composition of population in terms of economic status, formal education exposure and other relevant variables? I am convinced that assessment reports must be presented in terms of social context. This is another meaning for equity in assessment.

Comments on 'Equity in education and assessment'

Anil Kanjee

The issues/topics discussed here are critical issues currently being debated in South Africa, since, as you are probably aware, we are currently in a process of implementing a new education system. My comments on the paper on 'Equity in Education and Assessment' are thus based on our experiences in South Africa.

Understanding of equity

By definition, equity issues in assessment have to do with groups of people who have been disadvantaged or marginalized for reasons such as gender, language and/or race. The focus is on ensuring that no one

group of people is disadvantaged in the assessment process. In Gipps' paper there is recognition that, in practice, groups do differ and thus the preferred definition/understanding of equity is based on the equity of access rather than on the equity of outcomes approach. However, if not put in context, there is a real danger that these differences in performance between groups will be accepted as normal and identified as characteristics of that group. This is especially true with respect to minorities and/or disadvantaged groups.

Having said that, I think the emphasis on equity of access makes sense, as it is practical and can be achieved. However, the social, economic and political con-

Anil Kanjee

texts also need to be considered, since even if equality of opportunity/access exists, social, economic, political and/or cultural conditions may sometimes hinder the use of these opportunities. Thus, we can find situations where ample opportunity exists, but very few can make use of them. For example, an obvious situation to which the paper alludes is the area of mathematical sciences, which tends to have a very low representation of females. This indicates that the availability of access to opportunity is in itself not adequate to address equity problems.

In South Africa, disadvantaged groups, by definition, usually require more than the opportunity for equal access, and if additional resources are not provided access will not be utilized. A case in point is that of language. Providing equality of access to the curriculum, or funding opportunities for those who speak English as a second or third language (where English is the medium of instruction) did not achieve a higher representation of graduates from different minority groups in higher education, since the prerequisite skills or support required for success in higher education were not provided, thus leading to poor performance and higher failures. Another example: At one university, language and social support to help Black and female students adjust to the tertiary environment was a key component in a successful science programme when combined with the opportunity for greater access.

Another aspect is the issue of perception. Again, by definition, disadvantaged groups are those that have been discriminated against by the system. They tend to be suspicious of the system, including the education assessment system. In South Africa, assessment is a mechanism for social control to maintain and promote the status quo. It has done so, if one looks at the history and use of testing. For any equity initiative to succeed, this negative perception amongst the people for whom the initiatives are developed must be addressed. This is especially true for high-stakes assessment.

A point on definition and understanding of equity: I agree that equity issues must be put in context of the social, economic and political conditions. I would have liked the paper to have elaborated further on those aspects. With respect to assessment, Table 1, 'Curriculum and Assessment Questions in Relation to Equity', is critical, and I would add the following:

- One curricular question: What knowledge is taught?
- One assessment question: Who is responsible for assessment?

An interesting point that can be included here is the purpose of the assessment. That is, the use of the outcomes of assessment needs to be considered, especially in cases where these have some social value. For example, selection of medical doctors should not only focus on performance of students but also on the needs of communities and the intention of students to meet these needs. While difficult to assess, this nonetheless adds a social/community dimension to the whole assessment process and does address the issue of equity. This infers that access to education, generally subsidized by the state, is based not only on performance or ability, but also on community (state) needs, as well as the individual's community involvement. In South Africa in the late 1980s, historically disadvantaged institutions used a similar approach, that is, emphasis on community needs, student community involvement and so forth. It is clear that these facts should be included in addressing equity. The issue is not only equity for who, but equity for what.

Possible solutions

Having noted some of the difficulties, how can we, and should we, as assessment practitioners, address these issues? Certainly, changing social, economic, cultural and political conditions, as well as perceptions of groups are very difficult and complex issues. These can only be changed over a long

period of time and are great responsibilities of society in general. However, I do believe that we, as assessment practitioners, can ensure or promote some basic steps in that direction.

First, we need the commitment, will and support of the government/state. In South Africa our constitution highlights the issue of equity and notes the problems of delivery of resources. Also, legislation regarding the implementation of the new education system specifically notes issues of redress and equity. If the support of the state is not forthcoming, assessment practitioners/experts need to become greater activists or advocates in ensuring that this support is provided by the state. If it is available, we should make full use of this in the promotion of social change to highlight equity issues as it affects assessment. To this end, assessment practitioners should be involved in all aspects of education: curriculum, teacher training and so forth, and should also be involved at the conceptualization stages of any policy changes.

In addition, all our work, technical or theoretical, must have a stronger and greater societal focus. While I agree that this is not easy to do, I think it is necessary if we as assessment practitioners want to ensure that our work has the maximum intended impact.

Let me give an example from South Africa that is quite relevant for this discussion. Issues in South Africa are clearer, that is, black and white, literally and figuratively. In the field of psychology we did manage to function as professionals and as activists, and changed the whole field, that is, definitions of psychology, training programmes, curriculum and the professional organizations. In education, the intervention and participation of trade unions had a major impact on the introduction of outcomes-based education and on the introduction and development of the National Qualifications Framework that currently underpins the drive for a new education system. In the case of South Africa, it was important that assessment practitioners be

involved in all aspects of the education system: policy issues, curriculum development, teacher training and so on.

Definitions of groups

A major issue to revisit is how groups are defined. Besides the male/female groupings, I am not sure whether the use of ethnic and/or racial concepts are adequate in multicultural societies. I believe that socio-economic status or class-based classifications are more relevant. Another possible category/grouping to look at, especially with respect to minorities in western societies, is the effect of western influence or westernization of minority/disadvantaged groups. There certainly is a strong case to be made for this. For example, in Gipps' paper it would be interesting to know whether the Indian, Bangladeshi, Asian and African-Caribbean students were first-, second- or even third-generation citizens. I would like to think that this would certainly have some impact and cross-tabbing this information with socio-economic status also would provide interesting findings.

Conclusion

I conclude with four points, the first of which is that the issues of knowledge construction, knowledge use and assessment are critical aspects of equity. This area definitely needs more research. Further work on epistemology and assessment should prove extremely interesting.

Second, some of the ideas noted regarding practical issues of assessment in the Gipps paper are excellent and need to be followed up. For example, the idea of greater student or even parental involvement in the assessment process and the resulting impact on equity is certainly worthy of future research.

Third, the paper argues that differences among groups do exist and have to be considered. However, if these differ-

ences are not placed in context, I think that there is a real danger of accepting these differences as normal. I think the concept used in differential item functioning (DIF) research – that is, when comparisons between groups are made, the issue of ability levels of individuals within groups is critical.


Last, the idea of consequential validity, which I think should include equity requirements, is something that should be discussed as we attempt to bring equity issues into the realm of the technical requirements. I think it is an idea worth considering.





3. Education standards: current directions and implications for assessment

Howard T. Everson



Introduction



Education standards must play a defining role in education policy. Before we establish education standards, we must give careful thought to what it is we want our students to know and be able to do. What should students entering high school know in mathematics and literature? What should they be able to do with the skills acquired in history, language arts, and computer classes? Will they be prepared for the challenges of high school, college, the world of work? These are important questions for parents and others to ask. These questions are at the heart of the education standards movement in the United States. And, as we end the traditional school year in the United States, the themes running through this roundtable discussion – education standards, international comparisons, performance assessments and equity in education assessment – are especially timely.

It is also important to note at the outset that education standards are central to the work of the College Board. Although we are widely known as the spon-

sor of the Scholastic Assessment Tests (SATs) and the Advanced Placement tests, two assessment programmes considered by many as standards of academic excellence, the College Board has long been recognized as the organization that launched the debate over education standards in the United States by undertaking the Educational Equity Project in 1986. This effort resulted in the publication of the 'rainbow series' which detailed for the first time what students needed to know and be able to do to be prepared for college. More recently, Donald M. Stewart, the president of the College Board, underscored the importance of standards in an address to the 1993 College Board National Forum, noting that 'the most important role the Board continues to play is as a voluntary standard-setting organization for the world of education, typified by its leadership in facilitating the transition from high school to college (Stewart, 1993, p. 2).

As our notions of education standards change and mature, the connections (or disconnections, as some argue) between standards-led education reform and higher education admission require-

ments in the United States become more important. Questions about the nature and role of college admissions tests, such as SAT, become more central. Before focusing on the implications of education standards for college admissions testing, I wish first to set a larger context by providing a brief sketch of how education standards have been defined, and detail the process that educators have used to set standards and gauge student performance. The discussion will then turn to the connections between standards and higher education, and will conclude by presenting a research framework for advancing college admissions testing in ways that may be more compatible with and supportive of education standards.

■ ■ ■ □ **Overview of standards-based assessment**

The currents of standards-based education reform are shifting and changing in the United States (Linn, in press; Tucker and Coddling, 1998). Content and performance standards have been established in many academic domains, including mathematics, English, history, and the sciences. Individual states and the federal government have become key players in the implementation of education standards. Nearly every state has joined the movement to make standards central to education reform. Indeed, it is fair to say, following Tucker and Coddling (1998) that we are 'in the midst of a movement to use standards as the rallying principle for the improvement of academic achievement in the schools' (p. 41).

Defining standards

As in most policy-level debates, the discussions of national standards have created a good deal of confusion over just what we mean by a standard. The term has come to have different meanings for different communities of educators, parents, and policy-makers. This may be a good time to ask

what the term standard means in the education context. *Webster's Unabridged Dictionary* provides two general meanings for the word standard: (1) something established by general consent as a model or example to be followed; and/or (2) a definite level or degree of quality that is proper and adequate for a specific purpose. To those following the national debate, it is obvious that both definitions are widespread, contributing at times to a confusing buzz in education policy circles.

John Dewey, writing over six decades ago, pointed out in *Art as Experience* (Dewey, 1934) that a standard is a unit of measurement that functions symbolically and possesses none of the qualities of what it has measured. As such, a standard was simply a vehicle for describing a set of qualities. Dewey's sense of standard prevailed for many years in American education. SAT's familiar 200-to-800 scale is an example of a standard that has functioned symbolically as Dewey suggested. Today, as we begin asking the sharper question of what students scoring, say, 650 on the SAT scale know and can do, the utility of Dewey's definitional framework is less appealing.

Paul Barton, the director of ETS's Policy Information Center, in his excellent reference workbook, *National Standards for Education: What They Might Look Like* (1993), makes the point well when he notes that the term 'standards' has come to have several uses, among them:

- A clear statement of what students should know and be able to do at particular points in their schooling. The meaning here suggests that standards are statements of student expectations, for example, minimum performance standards in various domains and grade levels, particularly in reading, writing and mathematics.
- Performance levels that students should be able to attain or demonstrate. This idea begins to blend standards with assessment.
- Specification and definition of the necessary and desirable core of knowledge

Howard T. Everson

in a subject to be taught. When used in this context, the intent is to convey a sense of a canon or corpus of knowledge that should be taught and mastered.

Further evidence for the changing conception of the term is found in the increasingly widespread discussion of three different types of standards: content standards, performance standards and opportunity-to-learn standards. Again, these are important terms of reference in today's debates over standards.

Content standards refer to the narrative descriptions, the common understandings, of the desired outcomes in various subject areas. Performance standards, on the other hand, derive from defining and providing concrete examples of the level and quality of performance students must demonstrate to show mastery of the content outcomes. In other words, content standards refer to the 'what' of learning, and performance standards address questions of 'how well'. One can presume that performance standards will play a primary role in providing the frameworks for developing tests and assessments in a variety of subjects. In contrast, opportunity-to-learn standards are benchmarks for judging whether a state, a district, or a school has provided the resources, for example, a challenging curriculum, qualified teachers and so forth, needed to ensure that all students have the opportunity achieve high levels of performance. An example of both a content standard and a performance standard may be useful.

Delaware's content and performance standards in reading

The content standard in reading requires students to construct, examine and extend the meaning of literary-informative and technical texts through listening, reading and viewing. For example, to demonstrate their knowledge of this standard, fifth-graders must read a full-length passage from

a text and answer questions requiring both brief and detailed responses.

Based on how students' answers demonstrate their understanding of the passage, the performance standard indicates they 'meet or exceed' the standard if their answer:

- accurately summarizes the story or non-fiction sequence;
- identifies and discusses the characteristics (where appropriate) of the type of literature;
- identifies and explains technical elements of the language and how it was used in the story, giving supporting ideas to show a complete understanding of the selection;
- chooses facts or details relevant to the questions posed; and
- develops a justifiable and complete personal reaction to the selection, relating ideas in the story to personal experiences and to other reading, and evaluating the selection.

Developing standards

Since the late 1980s, raising standards in the major curriculum subjects has gained momentum in states and districts across the United States. Many observers credit President George Bush for his leadership when, in 1989, he invited the nation's governors to Charlottesville, Virginia, for the first national summit on education. Summit participants came away agreeing on the need for national education goals. A few months later, the Bush administration moved to create the National Education Goals Panel, an unofficial group of governors, administration officials and education policymakers who took the lead in monitoring the nation's progress toward the education goals. At about the same time, the United States Department of Education and other grant-making organizations began providing funds to a number of professional associations to begin developing national standards, discipline by discipline (Tucker and

Codding, 1998). In 1991, for example, the National Science Foundation's (NSF) State Systemic Initiatives funded standards development in math and science in twenty-five states. Similarly, the United States Department of Education's Office of Educational Research and Improvement funded the efforts of twenty-three states in 1993 as they worked to promote standards development in English/language arts, history, geography, civics, foreign languages, mathematics, science and the arts. The Department of Education's Dwight D. Eisenhower National Program for Mathematics and Science Education also funded the development of curriculum frameworks in math and science in fifteen states and in the District of Columbia. Similarly, beginning in about 1990, private foundations, for example MacArthur, the Pew Charitable Trusts and Carnegie, helped advance the process by convening standards-setting groups in collaborative organizations, such as the Council of Chief State School Officers.

Further impetus came in March 1994, when Congress passed the Goals 2000: Educate America Act, which provided funding to schools, communities and states to raise their education standards. Continuing the work of his predecessor, President Bill Clinton in October 1994 signed into law the Improving America's Schools Act (IASA), which renewed the Elementary and Secondary Education Act (ESEA) of 1965 and provided the authority for a \$10 billion appropriation in aid to states and localities. Passage of these legislative initiatives marked the beginning of the movement by local governments to develop and implement education standards designed to clarify what students ought to know and be able to do, and what content and performances ought to be valued by teachers. It also fuelled the movement for greater accountability in publicly funded education by providing benchmarks and standards of performance for the schools.

Looking back, we see a standards movement in the United States that was supported by state legislatures, local

communities, private foundations and the federal government. States and communities established their own standards, typically without direction from outside sources though often adapting the standards set by professional organizations such as the National Council of Teachers of Mathematics (NCTM), the National Council of Teachers of English (NCTE), the American Association for the Advancement of Science (AAAS), the National Academy of Sciences (NAS) or the New Standards Project (a collaborative of nineteen states, six urban districts, the Learning Research and Development Center of the University of Pittsburgh and the National Center on Education and the Economy).

Defining which standards are important and developing methods to measure them is difficult work, often involving a long and arduous process of consensus seeking, debate and clarification. In general, setting standards requires defining the 'essential' aspects of each subject and, in co-ordination with broad-based community groups, writing a rigorous core of priority standards that speak directly to the concerns of teachers and parents. Once standards are drafted, groups involving educators and citizens statewide develop plans to disseminate, review and implement them. Many states develop a monitoring strategy that tracks progress toward statewide adoption. Finally, to keep the purpose of standards in focus, states inform the public about the process and the definitions they use in developing standards.

Overall, between 1989 and 1997, a remarkably brief period, forty-nine states have begun to articulate challenging standards and revise curricula in the core academic content areas in response to Goals 2000 legislation. These standards, as we noted earlier, are designed to make explicit the content knowledge and cognitive skills students are expected to master during their elementary and secondary education. A first principle, and one that in many ways is yet to be tested, is that standards-based assessment is the key to education reform. The

rationale is straightforward. Standards-based assessments provide useful benchmarks for school systems, schools and teachers. Moreover, because standards are explicitly related to curriculum and instruction, the belief is that they will guide education practice. Finally, it is argued, standards-based assessments motivate and otherwise invigorate learning, teaching and schooling at all levels by giving both educators and students a concrete vision of what is to be achieved (Herman, 1992). Although a handful of states, including Vermont, Kentucky, Maine, Missouri and New Hampshire, have had modest success in implementing standards-based reforms, the view of many observers is that it is simply too soon to tell if the students are demonstrating stronger levels of achievement. Performance standards and assessments are needed to capture these outcomes.

Assessing performance

Standard-setting requires that we reach consensus on performance standards and, thereby, reach agreement on which performance proficiencies are appropriate in core subjects at different developmental levels. Performance standards gauge the degree to which students meet content standards. Clearly, this suggests that a number of elements must be considered in setting the standard, including the following: What qualifies as evidence that a standard is met? What will be the means or mode of assessing performance? How will we draw distinctions among proficiency levels? Thus, along with the articulation of content standards comes the need to develop assessment systems that will provide the evidence that students are meeting high standards of academic achievement.

The standards movement has encouraged many to think about changes in policy and practice with regard to assessment. Nancy Cole, the president of Educational Testing Services, in her presentation at the Conference on Partnerships for

Systemic Change in Mathematics, Science, and Technology Education (Cole, 1992), was eloquent in her characterization of this shift in education assessment when she said:

Standardized educational assessment in the United States has been an activity largely external to the instructional process. Although teachers have had various roles in planning test content and even preparing test questions, the driving force for standardized tests in the elementary and secondary schools has been policy-makers and governments. The driving purposes for testing in the schools have been for external monitoring and accountability ... Today, we are seeing a dramatic change in the driving force for assessment. Educators are trying to reclaim educational assessment and to shape it to serve purposes of teaching and learning first and foremost.

Reinforcing Cole's message, the NCTM standards have been widely touted as a framework for rethinking assessment standards and have been used as a point of departure for describing what is to be observed and measured in the process of understanding what students should know and be able to do in a variety of academic domains. The NCTM blueprint calls for measures of mathematical power, reasoning, problem solving, conceptual and procedural knowledge, and communication skills.

Another glimpse of what performance standards might look like is seen in the National Science Education Standards (1993). Like the NCTM approach, here we find standards for assessment practices in science that might possibly be generalized to other disciplines. They are paraphrased as follows:

- Assessment activities should focus on the content that is most important for students to learn.
- The form of assessment should be consistent with the valued content learning.
- Valid inferences about students' learning and achievement should be based on information from assessment activities.
- Assessment practices should be fair to all who are assessed.

- The assessment process should involve teachers and other professionals in the design, development, implementation and interpretation of assessment activities and the resulting information.
- The assessment process should give equal attention to assessment of the opportunity to learn and to assessment of student attainment.
- The design of the assessment process should be determined by the intended use of the resulting information.

Both the mathematics and the science standards emphatically stress the need for tests and assessments that promote learning and teaching, shaping education assessments to facilitate student achievement. If this view takes hold in the wider education community, it is easy to imagine a demand for changes in large-scale assessment systems such that they contain items and tasks that exemplify the standards in the relevant discipline, foster meaningful learning and have clear instructional value. Thus, as the national conversation shifts to a more direct and explicit discussion of expected student performance standards at the kindergarten through 12th-grade level, the implications of this movement for large-scale college admissions testing is likely to come into much sharper focus. Political pressure, no doubt, will require that higher education become an active partner with schools, likely through an alignment of admissions standards and criteria.



Connections to higher education

In the current reform climate, preparation for the world of work and higher education is highlighted in the discussion of standards. For the most part, colleges and universities have not had a direct role in establishing content and performance standards at the K-12 levels (Linn, 1994, in press). The National Governors' Association, in a report entitled *College Admission Standards and School Reform*, highlighted this absence not-

ing that school reformers have raised concerns about the need for colleges and universities to respond to the changes in curriculum, pedagogy and assessment that are taking place at the secondary school level. In particular, the reformers say that by clinging to conventional admission criteria, institutions of higher education are hampering schools' efforts to implement needed changes (p. vii). Further conversations centered on the connection between K-12 reform and college admissions took place in January 1994, when a group of eighteen representatives from public secondary schools and eighteen representatives from colleges and universities spent two days at the Harvard Graduate School of Education discussing this issue. Discussion, not unexpectedly, focused on what students are expected to know, how they demonstrate what they know and how best to create the connections between what is assessed in high school and what is valued in the college admissions process. They concluded by resolving to foster a closer collaboration among high schools, colleges and universities.

More recently, the American Association of Universities (AAU) Presidents' Committee on Undergraduate Education established a Task Force on K-16 Education, and charged the working group with exploring how colleges and universities could use the results of state assessment for admissions purposes. Acknowledging that colleges and universities have not played a major role in standards-based reforms at the K-12 level, the AAU task force will no doubt look into the question of how colleges and universities can influence the reform movement by specifying the knowledge, skills and attitudes toward learning they value when judging the qualities of their applicants for admission. From the perspective of the College Board, these are important conversations, highlighting the enduring need to create assessments that foster student success and facilitate the transition from high school to college.

Clearly, large-scale college admissions tests, such as SAT, need to be

responsive to the trend toward national standards. Test sponsors and test-makers are searching for ways to build on the performance standards that are expected to emerge from the work of various discipline-based groups like NCTM, the National Research Council on Science Standards and others. But what do these standards mean for tests such as SAT, a national test that has traditionally attempted to maintain a curriculum-neutral stance? How can we – and should we – re-engineer SAT to measure more directly student achievement in a standards-based educational environment? Will this effort fit the needs of our colleges and universities, who have traditionally relied on SAT to assist them in the admissions and selection process? These are but some of the questions facing us as we work to meet the challenges of standards-based assessment.

The SAT standard

If, as was suggested earlier, one adopts Dewey's view of an education standard as a unit of measurement with symbolic meaning, then clearly the SAT functions in American society as a standard. Indeed, many argue that SAT has served well as a national education standard for nearly half a century. This view crystallized in the early 1980s, when the then Secretary of Education, Terrell Bell, used SAT scores to create the United States Department of Education's Wall Chart, which served to rank-order the states and allow less sophisticated observers to draw inferences about the states' education performance. For Secretary Bell and others, the SAT continues to have a good deal of symbolic meaning and serves as an indicator of American education standards. Further evidence of the symbolic value of the SAT is found in the intense media attention that each year surrounds the release of the College Bound Seniors Report. Journalists from coast to coast report national trends and state-by-state rankings, and publish all manner of editorial comment and opinion. Their

doing so, one can argue, attests to the SAT as a powerful symbol of a national education standard in the Deweyan sense.

While arguing that higher education certainly needs to be more involved in the standards-based reform effort, Linn (1994, 1998) argues that there are a number of important complications when it comes to considering wholesale changes in admissions tests such as SAT. Among them is the fact that most colleges in the United States are simply not very selective, and in their struggle to attract students they may be unwilling to adopt the higher entrance requirements presumed by a standards-based approach to admissions. For the most selective institutions, Linn argues, there may be little of value in using results on standards-based assessments that do not provide sufficient rank-orderings of applicants. Linn (1998) also points to the possible inequitable uses of assessments that do not provide students with a 'second chance' to demonstrate their talents to college and university admissions committees. The inequities in opportunities to learn may also come to plague the standards-based assessment if they are used in the post-secondary admissions process.

Times may be changing, however. We now have different ideas about how children learn, and this has influenced the education reform movement. The expanding interest in reasoning and problem solving is well supported by research on learning. This shifting conception of learning has influenced our view of education measurement. As Nancy Cole (1992) reminds us, these changing views are finding their way up the policy-making ladder.

Underscoring the policy perspective, Marshall Smith (see Smith et al., 1990), the current Assistant Secretary of Education, remarked that 'the high stakes of SAT tests in this country have little effect on performance in school or on student learning in general, because SAT tests are designed to be largely independent of school curricula and outside preparation.' Like many others today, Secretary Smith has

been influenced by the emerging understanding of the nature of student learning and now advocates a movement away from reliance on curriculum-neutral, multiple-choice tests as indicators of meaningful learning. What we are seeing, then, is that the view of the SAT held by Secretary Bell, the SAT as a Deweyan standard, is changing. Moreover, as progress in education measurement continues, that is, as we learn more from our experiments in standards-based assessment design, this shift toward a more instructionally relevant view of testing and assessment, many believe, will become ascendant.

The role of SAT

A re-reading of *Beyond Prediction*, the report prepared for the College Board by the blue-ribbon Bok-Gardner Commission on New Possibilities for the Admissions Testing Program (1990), reassures us that SAT will remain responsive to the societal call for education reform. We would be remiss, however, if we did not also acknowledge that much has changed in the educational environment since that Bok-Gardner Commission drafted its report. In the past few years, there has been a rising tide of education reform accompanied by an active call for national standards in a number of academic areas. The standards-based reform has gained a good deal of momentum and has fueled an explosion of thinking, research, and development to address the calls for alternative forms of assessment. As we learn more about the teaching and learning process from psychology, education measurement and instructional science, the enthusiasm for these new and alternative forms of assessment will likely grow, and challenges to the validity and relevance of the more traditional forms of assessment, like the curriculum-neutral SAT, are likely to increase. Thus, along with the broadening of the name to Scholastic Assessment Tests, so as to more accurately reflect the increasingly wider array of assessments that comprise the SAT, the role for these assess-

ments in the future will be broadened as well.

For SAT to remain relevant and valued in what is quickly becoming a new and profoundly different educational environment, the College Board must remain vigilant and continue to develop and redevelop this assessment programme in ways that not only retain its considerable strengths, but also capitalize on what we have learned about standards-based assessment from the Advanced Placement and Pacesetter programs. SAT, for example, may have to move beyond its traditional role of sorting and selecting and strive to better serve the purposes of teaching and learning. By doing so, SAT will be better positioned to remain the premier instrument for use in the college admissions process and retain its value as an education standard in the United States.

Changing may mean creating a SAT that in the future will emphasize the measurement of reasoning, problem-solving and critical thinking in the context of well-defined subjects, such as history, mathematics and geography to name a few. Student performance under this scheme would no longer be reported solely on the familiar SAT scale of 200 to 800, but would also be communicated in terms of the performance standards described by the various discipline-based groups currently at work on defining the national standards. This may mean, for example, reporting beginning, developing or advanced proficiency levels for each student and in each of the subjects in the test battery. The challenge will be to create an instrument, or set of instruments, that would not only provide colleges with information needed to select applicants and place students, but also ensure that the SAT is more closely aligned with the content and performance standards emerging across disciplines.

Others argue it is possible that changes to SAT could very well take another path, one that retains SAT's current emphasis on the measurement of critical reading, verbal reasoning, mathematics and writing,

but is broadened to link aptitude and achievement measures. While not ignoring the movement toward national standards, this perspective offers a framework for SAT of the future that, in general, is less likely to be constrained by discipline-based standards, suggesting instead an evolutionary approach that would build on the SAT I Reasoning tests and the SAT II Subject tests in writing and mathematics, thus making SAT more useful to colleges and universities as they work to place students in appropriate courses.

These two somewhat different approaches will guide the early phases of research and development of assessment prototypes for SAT. From a research and development perspective, there is much to be learned from a full-scale exploration of both approaches. We need to know, for example, how far we can push the boundaries of reasoning tests in an effort to link measures of aptitude and achievement. Pursuing test development models and prototypes that attempt to broaden the current SAT, as well as those that deliberately address standards-based reform models, allows us to explore possible new and varied roles for SAT, and may also lead to a long sought-after rapprochement of aptitude and achievement testing.

This exploration will require that we find ways to examine the implications of these overlapping and somewhat competing approaches. If, for example, SAT is developed as a standards-based measure, how well will it serve its traditional role in the college admissions and selection process? How will we address the concerns of equity raised by Linn (1988) and others? On the other hand, if future versions of the SAT, whether they be computer-based or paper-and-pencil tests, are created, whole cloth, as expanded and broadened variations of the traditional measures of verbal and mathematical reasoning, will they remain relevant and useful in a standards-based education reform environment? Will the SAT of the future help to clarify the signals we send to our nation's students and

schools about education standards and what knowledge is worth knowing (Kirst, 1998)? These are important issues and challenges, and we have to adopt criteria, both psychometric and educational, to guide our research and development efforts and investments.

Research and development criteria

Fortunately, criteria for creating and evaluating new assessments are beginning to appear in the education measurement literature. Again, researchers such as Robert Linn (see Linn, 1991; 1998) and others have contributed to our thinking in this area. They have articulated an expanded set of criteria for evaluating the quality and implications of a variety of alternative assessment approaches. The criteria they have developed, along with the psychometric values that have long supported SAT, lend themselves to our effort to redevelop large-scale admissions tests. These criteria include the following:

- *Consequences.* How the results derived from each of the various approaches to the future SAT would be used, as well as what their intended and unintended effects may be, can and should be assessed at the earliest stages of the research and development effort. Test development models that allow for an expanding role for SAT are attempts at addressing this criterion. By deliberately starting out to develop alternative models and approaches, we are in a stronger position later on to assess their consequences, one against the other.
- *Fairness.* Here the focus must be on determining the equity implications of the various approaches. Will, for example, all students have an opportunity to learn what is assessed if SAT becomes a standards-based assessment? Does one approach present greater advantages to one group of students over another? What about issues of coaching? Is a standards-based test more susceptible to

short-term, intensive coaching? Will computer delivery foster or hinder equity in assessment? These questions and concerns can all be grouped under the criterion of fairness.

- *Transferability and generalizability.* This criterion suggests that we need to examine carefully the question of whether the results derived from the various models, and prototypes support accurate generalizations about student abilities and potential. How many and what types of tasks will be needed to ensure that the assessments have acceptable levels of generalizability? What level of content coverage is appropriate?
- *Cognitive complexity.* The issue here is to examine whether the approaches are equally effective in assessing higher-order thinking skills. Can we get closer to the goal of measuring problem-solving, reasoning, and conceptual and procedural knowledge using one approach or the other? Our challenge here is to develop sound theories of domains we are testing, along with reasonable models of how the learner progresses from novice to expert in those domains.
- *Content quality and coverage.* The competing approaches need to be evaluated in terms of how well they represent the content standards, and we need to ask whether they are consistent with our best understanding of the important aspects of the disciplines valued by education reformers. This implies that we have to work closely with professional groups and disciplined-based organizations to ensure that high levels of content coverage are achieved.
- *Meaningfulness.* The issue here is whether the suggested approaches provide equal vehicles for providing assessments that assure educators, policy-makers and others that students are engaged in meaningful problems and worthwhile educational experiences. Again, the educational value of the various approaches for the future SAT will have to be addressed.

- *Cost and efficiency.* Attention needs to be given to efficient testing and scoring methods. The latter have direct implications for scale choice, as well as test format and test delivery methods. Given our work with SAT and other large-scale testing programmes, the College Board has a decided advantage in this regard. Model choice and prototype development, along with issues of scale choice, need to be guided by these criteria as well.

Conclusion

The research and development challenges posed by the emergence of the standards-based assessment movement are formidable. Viewed separately, the two developmental approaches for admissions testing outlined earlier – the evolutionary and standards-based perspectives – can be seen as charting somewhat different directions for the next generation of college admissions tests. Keeping in mind that the current uncertainties of today's education assessment environment are likely to yield to the pressures of the standards-based reform efforts, the implications of these two approaches remain, at least for now, somewhat unclear. From a research perspective, our aim at this stage is to generate and shape the discussion of issues of feasibility, calibration, fairness and relevancy as our research and development agenda unfolds. Emphasizing the need for a strong research base, Donald Stewart reminds us that 'we must, when dealing with the lives of our young, advance cautiously upon the back of solid and extensive research, prudently avoiding the rushed or intemperate reform' (Stewart, 1994, p. 12). If, in the Deweyan sense, a standard is a means for describing a set of qualities, then the standard set by the College Board's assessment programmes, as exemplified in SAT, Advanced Placement and Pacesetter, will undoubtedly endure.

References

- BARTON, P. 1993. *National Standards for Education: What They Might Look Like*. Princeton, N.J., Educational Testing Service.
- COMMISSION ON NEW POSSIBILITIES FOR THE ADMISSIONS TESTING PROGRAMME. 1990. *Beyond Prediction*. New York, The College Board.
- COLE, N. 1992. *Changing Assessment Practice in Mathematics Education: Reclaiming Assessment for Teaching and Learning*. (Paper presented at the Conference on Partnerships for Systemic Change in Mathematics, Science, and Technology Education, Washington, D.C.)
- DEWEY, J. 1934. *Art as Experience*. New York, Minton, Balch and Co.
- HERMAN, J. L. 1992. *Accountability and Alternative Assessment: Research and Development Issues*. Los Angeles, Calif., National Center for Research on Evaluation, Standards, and Student Testing. (CSE Technical Report 348.)
- HOUGHTON, M. J. 1993. *College Admission Standards and School Reform: Toward a Partnership in Education*. Washington, D.C., National Governor's Association.
- KIRST, M. W. 1998. Bridging the Remediation Gap. *Education Week*, Vol. 18, No. 1, p. 76.
- LINN, R. L. 1991. *Alternative Forms of Assessment: Implications for Measurement*. (Paper presented at a symposium entitled 'Assessing Students' Mathematical Understanding: Issues that Underlie the Development and Use of Alternative Forms of Assessment' at the annual meeting of the American Educational Research Association, Chicago, Ill.)
- . 1994. The Education Reform Agenda: Assessment, Standards, and the SAT. *The College Board Review*, No. 172, pp. 22–5, 30.
- . 1996. *National Science Education Standards*. Washington, D.C. National Research Council, National Academy Press.
- . 1998. Implications of Standards-based Reform for Admissions Testing. In: S. Messick (ed.), *Assessment in Higher Education*, pp. 73–90. Mahwah, N.J., Lawrence Erlbaum Assoc.
- SMITH, M.; O'DAY, J.; COHEN, D. 1990. National Curriculum American Style: *American Educator*, Vol. 14, No. 4, pp. 10–17.
- STEWART, D. M. 1994. *Holding onto Norms in a Sea of Criteria*. (Paper presented at the conference Beyond Goals 2000: The Future of National Standards and Assessments in American Education, the Brookings Institution, Washington, D.C.)
- TUCKER, M. S.; CODDING, J. B. 1998. *Standards for our Schools*. San Francisco, Calif., Jossey-Bass.

Comments on 'Education standards'

Gordon Ambach

I have long been interested in testing and especially in the issues of international comparisons and what we can learn from

the world's virtual laboratory. I am not a psychometrician. My 'day job' is as a lobbyist. I am a political junkie. My job is to

Comments on 'Education standards'

round up one vote more than half for educational issues. In this sense, you can say I am on the side of those who are concerned with educational policy.

Everson provided a good background on terms and presented examples of the use of standards and assessments. His paper raised some important questions about how we tackle potential issues or conflicts between state or local standards and college admissions tests such as SAT or the American College Admissions Test (ACT).

Two major questions that I will focus on in my response are:

- How do policy-makers use standards and tests?
- How do standards and tests influence policy-makers?

The tenet

According to pedagogical theory, by making more explicit our objectives for learning, by gaining greater consensus and by holding students and schools accountable for success on these standards, achievement of all students will increase. Of course, achievement must be measured.

How it plays

In the United States and in much of the world, we are roaring along in a standards and assessment feeding frenzy: classroom testing; school testing; school district testing; state, national and international testing. There is more testing than ever before.

Politics also comes into the picture. To illustrate, I will share a personal story. In 1989 I became the United States representative to IEA. The first IEA meeting I attended was in Beijing in 1990. I had three missions:

- to try to persuade IEA to go ahead with TIMSS as a math and science study;
- to try to accommodate the United States' interest to add Grade 12 to the study in addition to Grades 4 and 8; and

- to discuss whether there should be performance assessment built into TIMSS.

The role of politics

National repercussions from TIMSS

No one could have anticipated the policy or political consequences of TIMSS in the United States:

1. On 4 June, President Clinton announced the United States Grade 4 test results from the Rose Garden at the White House.
2. The most frequent use of TIMSS data in the United States, especially results for Grade 12, was to demean public school results as part of an argument for financial support of private schools.
3. A cluster of suburban schools around Chicago organized to take TIMSS to demonstrate to businesses that they were world-class.

These consequences, though not predictable, are very important for the future of standards and assessments. We must understand the impact of standards and assessment strategy on education in the long term, but we cannot underestimate the role the political agenda plays in their development.

Policy issues

The following list of policy issues illustrates what is important politically in the United States:

- to increase overall student achievement, especially in math and science;
- to accomplish this goal by raising the performance of the lowest 75 per cent to 'close the gap';
- to accomplish the first two goals by centralizing or decentralizing (a controversial issue in the United States);
- to accomplish the first two goals by giving public money to private schools; and

- to accomplish the first two goals by using resources more equitably.

It is not difficult to see the role of standards and assessments in these issues.

Setting the standards and getting agreement on them is not easy. Standards are set at three levels: local, state and national. The situation becomes even more difficult when different types of standard – content, performance and opportunity to learn, need to be considered. It is clear that these standards should ‘fit’ together. However, the conservatives in the national government prefer to deal only with content and performance standards, leaving opportunity-to-learn standards to the local level. Since the United States has 15,000 school districts, it is easy to see how wildly these opportunity-to-learn standards would fluctuate if they were set at the local level. In addition, there are voluntary national standards and state standards. How do they differ and which should prevail?

Who wants what?

The public and government officials want results and not descriptions of inputs.

- For business, the bottom line is financial.
- For education, the bottom line is student achievement. States are competing with each other.

To complicate matters, as Everson points out, there are different views and uses of ‘standards’. They range from ‘minimum competency’, to ‘consensus on objectives’ to ‘best practice’. In addition, standards are only as useful as the assessments that measure them and linking standards to new assessments is a big issue in the United States.

The unique political context for standards in the United States

While there are limits to the role of the federal government, it puts pressure on states and localities by specifying national goals.

The federal government cannot set unequivocal national standards, but it can encourage states and localities to set goals or standards.

States can now require local school districts to meet standards and school district authorities to put measures of accountability on individual schools. These pressures have a great deal of impact on education.

National importance complicates assessment of standards

The importance the federal government attaches to the assessment of standards is indicated by its funding of the tests known as the National Assessment of Educational Progress (NAEP). NAEP, an on-going assessment of achievement in various subjects at three grade levels, has become known as ‘The Nation’s Report Card’. Then there is a national voluntary test. As significant as NAEP and the national voluntary test are, they further complicate an already complicated scenario. For example:

- There are differences between national voluntary standards and state standards.
- There are differences between NAEP assessment frameworks and state and local assessment frameworks.

It is difficult to explain the differences in the tests to the public who think ‘math assessment is math assessment’. To the educators, the tests measure different things.

The other issue that confounds the scenario is the question of how to merge standards for students going on to post-secondary study with standards for students who will enter the world of work following high school.

How standards and assessment impact on educational policy

From my perspective some of the consequences are:

- Battles over cultural values in standards,



for example, arithmetic versus algebra. There is intense public discussion about education.

- Scorecards and rank-order are used. This information provides motivation for reform.
- There is some use of analogies for pedagogical decisions.
- Education governance and power are changing: state power over school districts, and school districts' power over schools, with respect to accountability reports on 'external standards'.
- There is public/political influence over professional practices.

The international perspective

Now I turn to standards and assessment and education policy from an international perspective.

First, I will look at international comparisons/world class standards. Where do we stand competitively with others? In the United States there is increasing interest in international studies and comparisons with other nations: an education score card, benchmarking to other nations. How did we do on TIMSS compared to other countries?

Second, I will examine how results are used in political strategy.

- They are used to bash public schools and to support school vouchers, charter schools and other changes.
- They are used as ammunition against the status quo even though the data do not suggest solutions.
- In science and mathematics there is more interest in background variables related to performance. Hundreds of thousands of videotapes were distributed to teachers across the country showing how teachers in other countries teach math and science. This was an unpredictable consequence of TIMSS. The curriculum also came into question: breadth or depth?

Questions for this symposium

Having given the perspective of the education lobbyist in my response, I need to ask several questions from that perspective as well.

- Are there world-class standards or are we looking for the lowest common denominators or agreements among test developers?
- Is it possible to agree on world-class standards or achievement levels that are criteria-based and not curved?
- Can we explain or justify the differences in the assessments: international to state to local?

Questions for IAEA

- Can IAEA help set common standards across tests?
- Can IAEA help with the problem of efficiency of linking test results and reports?
- Can IAEA help create a long-term system for collecting assessment data that can be analyzed using data from UNESCO and achievement test results from studies by IEA, the Organisation for Economic Co-operation and Development (OECD), etc.
- Can IAEA help to build an international capacity for testing and processing and assist developing nations to develop resources for educational assessments?

Politics, money and education

Finally, returning to the role of politics in educational standards, assessment and policy, we need to bear in mind that politicians like standards and assessments because they are inexpensive. It's a lot less expensive to set standards and assessments than it is to finance programmes that provide oppor-

tunities. Perhaps that is not true in your country, but it is here. It is worth thinking

about when discussing the role of standards and assessments in educational policy.

Comments on 'Education standards'

Anton Luijten

believe that Everson is somewhat optimistic about the role of assessment in education reform. Public concern over quality control is often associated with a concern about changing the structure of education, comparing the progressive, pupil-centred approach with the traditional, more rigorous, teacher-oriented approach. It is not 'politically correct' to talk too loudly about monitoring education and quality control during a period of reform of the education system. Testing and examining bodies are the gatekeepers in the education field; but during a period of renewal and restructuring, the gates are open to everybody and everything: a free entrance for new ideas. In the euphoria of the renewal of the education structure there has to be a 'body' that keeps its feet firmly on the ground.

■ ■ **Concepts of standards** ■

Perhaps more strongly than Everson suggested, different points of view about standards hamper discussions about standards. There are three possible connotations of the concept of standards.

Political and social expectations

From a political perspective, the progressive wing will always emphasize the point that

performance standards are out of date and, at the same time, emphasize the responsibility of the government in the national control of education. In this view education has to focus on the needs of a constantly changing society. As a result, standards are constantly subject to changes and therefore hard to establish or maintain. Conservatives take the opposite view. There are certain age-old education goals connected to a certain gold standard. In this view it seems that since the time of Aristotle, the level of education has worsened. At the same time, the liberal view on education is to give more autonomy and responsibility to individual schools. Consequently, the context of the discussion about standards will depend on the prevailing view. The political and social expectations create the context in which the discussion about standards takes place. They form the context, not a practical framework.

Standards as education goals or objectives

Since the late 1960s, much work has been done to define domains, goals and objectives, resulting in the setting of attainment targets (or even national criteria) by recognized organizations and institutes. During the past decade, too much of this important work has been left to 'the Establishment', and the interests of pupils, parents, trade and industry have been neglected. For example, during the past five

years, the Ministry of Education in the Netherlands has tended to appoint independent committees and independent experts to set new programmes and attainment targets. Institutes for curriculum development and assessment are increasingly in the position of giving assistance to these committees, instead of applying their specific expertise in this field directly.

Standards as measures of education achievement

Results on standardized tests and examinations make an important contribution in attempting to attain quality control in education. Opponents of this quantitative approach of defining education standards will be alert to the danger that education goals will be made subordinate to, and dependent on, the 'measurability' of the

objectives. Of course, 'we don't learn for the sake of learning, but for life', but it is very helpful in life to have mastered education objectives.

Summary

The political concept of standards is not very helpful in attempting to formulate or to set standards in education. It only creates the framework for the setting of standards. The two pillars in the framework of standards have to be built by both curriculum and measurement specialists. It is unacceptable to leave the defining of standards to one institution, one committee or one single board. The acceptability of standards requires the involvement of institutes and committees representing different expertise and different commitments.

Comments on 'Education standards'

Kent McGuire

Introduction

Everson did us a service, by laying out the different concepts of what standards are, because they do mean different things to different people. Many times standards are in the eyes of the beholder. One very real issue in the United States today is: Whose standards? Who in fact makes them and who decides how are they to be used? Are they to be used to assist students or are they to be used as a means of controlling access to opportunity?

Several quick observations follow the idea of being more explicit about

what students ought to know and be able to do at a particular time is a generally positive development in the United States. In my view, the idea of standards is really fundamental in trying to have a clear, honest, open, public debate about the role of public schools in a modern democracy.

Impact on education

Gordon Ambach reminded us that the issue of standards has opened up quite an emotional debate about what the role of the schools is and what youngsters ought to know and be able to do at a particular time.

Kent McGuire

Sometimes, it appears that the opportunity to meet standards for some youngsters may come at the expense of others. Therefore, I don't think we have begun constructively to utilize these possibilities for standards nearly as much as we should, because the debate about them has been short-lived in the race to implement them. There is a lot of effort in curriculum reform and professional development efforts in assessment reform all designed to enact standards. This has been the kind of policy response in which assessment emerges as the dominant reform strategy.

I also agree with Ambach that it is cheaper, relatively speaking, to argue for higher standards and come in with an assessment or accountability system and strategy to see whether or not they have been met. It is a lot more difficult to conceptualize, implement and support education efforts necessary to meet the goal of having, say, 75 per cent of students achieve the standards. So the accountability system, at least in the United States, is way out in front of building the necessary capacities to improve the teaching and learning process. There is every reason to expect that things will get a lot worse before they get better, with the return to standards-based reforms, or something called standards-based reform. The new assessments tend to raise the bar. They elicit institutional responses, but the changes in teaching and learning lag seriously behind the assessments.

Everson's reference to Pacesetter is indicative of a widely held assumption about what is required if standards are to have the desired impact in schools. To what extent should we view Pacesetter as an integrated programme of instruction, assessment and professional development? We see a large array of models, instruction systems and designs, along with legislation intending to foster the expansion of these integrated programmes at the school level. An interesting question is whether we have enough evidence about the relationship between these models and student achievement to warrant their implementation.

Standards and higher education

Moving to Everson's discussion of SAT and higher education, I think another interesting question is whether there is an analogy for standards in higher education. There is, of course, the concern regarding whether standards in Grades K through 12 meet the criteria for admission to college. So changes to SAT are not unimportant. Another serious issue is the question of teacher qualification in elementary and secondary schools, and the standards to which *they* were held for graduation from teacher-education programmes. The effort to move people through the public schools is not inconsequential. Everson mentioned Equity 2000, which focuses on that transition. We can take credit for increasing the number of people who enter higher education, but we should be uncomfortable that the completion rates are more dismal in higher education than they are for the public schools. Some 40 per cent or less of students who enter higher education actually complete their studies. This leads us to the question of the quality of the undergraduate experience. What is the role of higher education in preparing teachers who have enough content knowledge and expertise to impart standards-based education?

Impact on assessment policy

On the question of the use of assessment for policy, I want to reiterate Ambach's remarks that at least in the United States, there has been an amazing national response to TIMSS. I have just started to learn the policy agenda of my colleagues in Washington. There is no question that TIMSS has been used as a device for managing and setting policies. The data from TIMSS have been important in focusing on how to make a coherent series of federal policy initiatives to provide a co-ordinated

response to what the TIMSS data suggest. Large-scale assessment data can and do influence policy. At the same time, as I mentioned before, it has been difficult to align the assessment results with the necessary capacity-building work. Therefore, the policy response you get may not always be the one you would hope for. I think Ambach makes the point well when he talks about the use of 'report cards' for accountability purposes instead of capacity-building.



Challenging the assessors

In closing, I hope that the measurement community will spend some time and energy on the kind of improved assessments and assessment strategies at the same time that it struggles with technical issues around large-scale assessment: assessment that would provide useful feedback for teachers,

parents and students. Feedback would provide some guidance on how to improve teaching and learning as we move to standards-based reform. Ranking students and countries is not enough.

While on the one hand assessments may go a long way to confirm our suspicion that some groups of youngsters do less well than others and that some have had less access to rich learning opportunities, assessments do not provide suggestions for remediation.

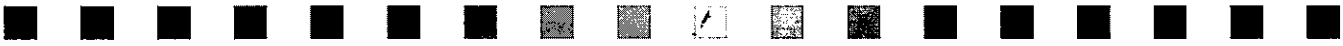
Messick (Chapter 1) raised the question about the potential implications that various technologies might play. One promising window on the future is that perhaps we can harness the prevailing and emerging technologies to provide better assessments and to determine how to provide this information to more people who need to know in order to effect change.





4. Performance assessment

Barry McGaw



What counts as performance?

The term 'performance assessment' is used in different ways, ranging from a very broad definition that includes any assessment requiring construction of a response to a restricted definition including only physical performances in real-life settings.

Where there is a strong tradition of multiple-choice tests, performance assessment is sometimes taken to include any assessment requiring students to construct or generate a response. From this perspective short-answer questions and essays are 'performance assessments'. The defining characteristic is construction of a response.

Under another use, the term 'performance assessment' refers to 'authentic assessment'. In this case, the assessment is focused on real-world, often on-the-job, performance. Here, the defining characteristic is context.

A third option is to restrict the use of the term 'performance assessment' to assessment of activities that involve performance in the everyday sense of the word. This would cover activities such as dancing, playing a musical instrument, performing

gymnastic routines, using scientific apparatus, speaking and so on. Unless the performance can be recorded for later analysis, assessment will rely on observation and judgement of an essentially ephemeral activity as it occurs.

A fourth option is to extend the third conception of 'performance assessment' to include sustained performance over a period of time to develop a product that would then be the primary focus in the assessment. Examples would include creation of a piece of art and design, and construction of a piece of equipment.

In this paper, I take the position that to call all assessment other than multiple-choice tests performance assessment is to adopt too broad a definition, while to restrict the definition to cover only real-world, on-the-job performance is too narrow. I choose instead to consider as performance assessment, the assessment of both actual performances on a single occasion and the products of sustained performances, where the performances themselves are not readily observable.

The legitimacy of performance assessment can scarcely be in doubt. The challenge is to devise valid and reliable methods of performance assessment.



Monitoring education systems

Need for curriculum fidelity in assessment

Many education systems currently monitor their performance in terms of the achievements of their students. The framework for such monitoring is provided by the curriculum and should reflect faithfully the breadth of the curriculum. Where the curriculum specifies desired outcomes in terms of performance, performance assessment is required. In systems with curriculum responsibility devolved to schools, the framework needs to be developed from an analysis of the curriculum policies and practices of the schools. In those with system-level responsibility, there are system-level curriculum documents to provide the framework.

The Australian curriculum profiles set out, in a sequence of eight levels, the progression in learning outcomes expected to be achieved by students during their ten years of compulsory schooling. For each outcome, sample indicators of achievement are provided together with annotated samples of student work that demonstrate achievement of one or more outcomes at the particular level.

The curriculum profile for English has three strands: speaking and listening; reading and viewing; and writing, each of which is subdivided into the same four strand organizers: texts; contextual understanding; linguistic structures and features; and strategies. The outcome statements for two of these four strand organizers in the Speaking and Listening Strand are shown in Table 1.

The sample indicators of achievement for Level 5 in the Linguistic Structures and Features strand organizer and the Strategies strand organizer in the Speaking and Listening strand are shown in Table 2.

With Speaking and Listening, it is clear that the outcomes and indicators include many for which paper-and-pencil tests could provide no useful assessment. With listening, written responses to tape recorded or other controlled speech could be used, but with speaking there is no substitute for performance assessment.

Performance data are more difficult to collect than are paper-and-pencil assessments. Systems that monitor by testing all students in a cohort typically focus on a restricted range of student learning that can be assessed with paper-and-pencil tests, often with a further restriction that responses can be machine-scored. In the New South Wales Basic Skills Testing Program in Australia, for example, where all students at Grades 3 and 5 are tested, only two aspects of literacy are monitored, viz. reading and language. Assessment of language is restricted to tasks such as editing, using items requiring selection of missing words or location of errors. There is no attempt to monitor students' achievements in those parts of the English curriculum concerned with speaking and listening.

Systems that monitor by testing only a sample of students rather than the whole cohort can obtain more extensive and richer information. The total dataset with which they have to deal is so much smaller that they can, for example, deal with open-ended responses requiring judgemental scoring and not restrict themselves to questions with machine-scorable responses. The Western Australian Monitoring Standards in Education programme involves annual surveys of different curriculum areas using only samples of students at Grades 3, 7 and 10. Assessment of all students, for reporting to parents and review within schools, can be achieved with an additional set of materials equated to those used in the statewide sample monitoring, published in the following year. When student achievement in English is assessed, performance in speaking and listening is included.



Table 1. Outcome statements for two strand organizers in one strand in English

| Level | Strand: Speaking and Listening | |
|-------|--|--|
| | Strand organizer: Linguistic structures and features | Strand organizer: Strategies |
| 1 | Draws on implicit knowledge of the linguistic structures and features of own variety of English when expressing ideas and information and interpreting spoken texts. | Monitors communication of self and others. |
| 2 | Experiments with different linguistic structures and features for expressing and interpreting ideas and information. | Speaks and listens in ways that assist communication with others. |
| 3 | Usually uses linguistic structures and features of spoken language appropriately for expressing and interpreting ideas and information. | Reflects on own approach to communication and the ways in which others interact. |
| 4 | Controls most linguistic structures and features of spoken language for interpreting meaning and developing and presenting ideas and information in familiar situations. | Assists and monitors the communication patterns of self and others. |
| 5 | Discusses and experiments with some linguistic structures and features that enable speakers to influence audiences. | Listens strategically and systematically records spoken information. |
| 6 | Experiments with knowledge of linguistic structures and features, and draws on this knowledge to explain how speakers influence audiences. | Critically evaluates others' spoken texts and uses this knowledge to reflect on and improve own. |
| 7 | Uses awareness of differences between spoken and written language to construct own spoken texts in structured, formal situations. | Uses a range of strategies to present spoken texts in formal situations. |
| 8 | Analyses how linguistic structures and features affect interpretations of spoken texts, especially in the construction of tone, style and point of view. | Uses listening strategies which enable detailed critical evaluation of texts with complex levels of meaning. |

Source: *English – A Curriculum Profile for Australian Schools*, Carleton, Australia, Curriculum Corporation, 1994.

Table 2. Level 5 indicators of outcomes for English

| <i>Strand organizer: Linguistic structures and features</i> | <i>Strand organizer: Strategies</i> |
|---|--|
| <p><i>Outcome:</i> Discusses and experiments with some linguistic structures and features that enable speakers to influence audiences.</p> | <p><i>Outcome:</i> Listens strategically and systematically records spoken information.</p> |
| <p><i>Evident when students for example:</i></p> <ul style="list-style-type: none"> • Observe and discuss the way that voice and body language affect audiences and can be used to enhance meaning and influence interpretation (the way gestures, posture, facial expression, tone of voice, pace of speaking may engage the audience's interest). • Note aspects of language use, such as vocabulary, rhythm, similes, which enhance particular spoken texts. • Discuss and experiment with the effect of intonation on meaning (say the same word, phrase or sentence in different ways to convey regret, anger, annoyance, humour). • ... | <p><i>Evident when students for example:</i></p> <ul style="list-style-type: none"> • Prepare for listening (take pen and notebook or laptop computer to the viewing of an information video or a talk by a guest speaker). • Note cues such as change of pace and particular words which indicate a new or important point is about to be made. • Develop and use a personal abbreviation system to record information quickly. • ... |

Source: *English – A Curriculum Profile for Australian Schools*, Carleton, Australia, Curriculum Corporation, 1994.

Assessing speaking performance in English

The 1995 Western Australian monitoring in English covered reading, viewing, writing, and speaking and listening (Cook et al., 1997). Oral language had first been assessed in a pilot study as part of the assessment programme in 1991. Speaking was assessed in group work and in individual presentation. Students worked in groups of three to five, with the performances of the whole group and one particular student in the group to be observed and assessed. Individual students were also assessed in individual presentations of two types, one a persuasive, expository presentation of the ideas from the group discussion, the other a narrative presentation of a familiar fairy tale prepared with assistance from the group.

The teachers were trained for this assessment task with a videotape of samples of group and individual student per-

formances. There were three forms of material on the tape:

- samples of group and student performances together with assessments of the performances and a discussion of the assessments by one of the test developers at the Australian Council for Educational Research (ACER);
- samples of group and student performances that the teacher was required to score while watching before checking in the training manual to see how the 'experts' had scored these performances; and
- samples of group and student work that the teacher ('markers') was required to score without support from any further explanatory material, with those assessments to be submitted for use later in studying marker reliability.

The results of the assessment are not the focus of attention here, though it is interesting to note that students performed

Table 3. Range of teacher ratings of speaking performance (0-7 scale)

| Teacher groups | Grade 7 Whole Group | Grade 7 Nominated student | Grade 10 Individual presentation | Grade 3 Individual presentation |
|----------------|---------------------------|---------------------------------|--|---------------------------------------|
| Grade 3 | 1-7 | 1-7 | 3-7 | 0-4 |
| Grade 7 | 1-7 | 1-7 | 1-7 | 0-5 |
| Grade 10 | 2-7 | 2-7 | 4-7 | 1-5 |

much better on the narrative than on the expository speaking task (the Grade 7 narrative mean being above the Grade 10 expository mean, for example). Here, the important question is about the reliability of teacher judgements of live and ephemeral classroom speaking performances gathered as part of the monitoring programme. The assessments of the final segment of speaking performance on the train-

ing tape provide the data to answer this question.

The data for the reliability study came from more than 200 classroom teachers (79 Grade 3, 75 Grade 7 and 64 Grade 10 teachers) (Mendelovits, 1997). All teachers assessed the same four speaking performances on the videotape:

- a whole group discussion by a group of Grade 7 students;

Figure 1. Mean ratings and spread (two standard deviations) for teacher ratings of speaking

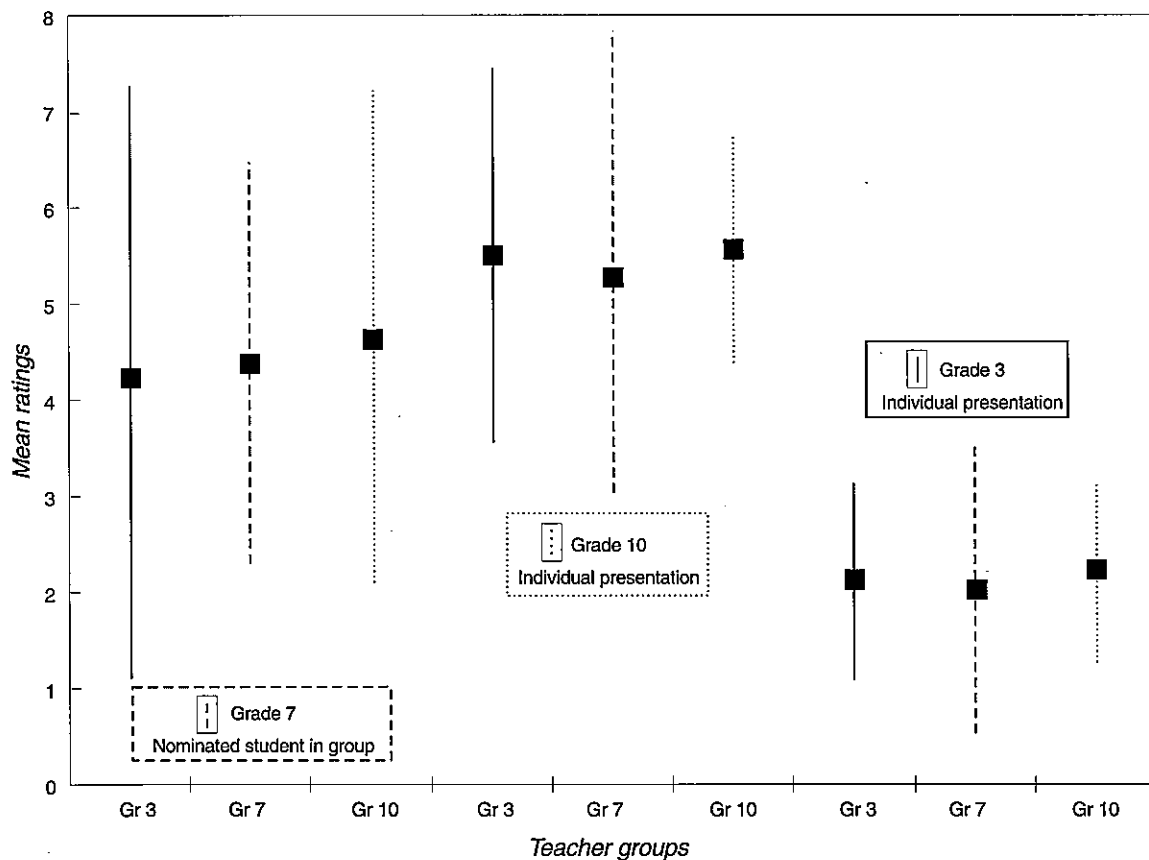
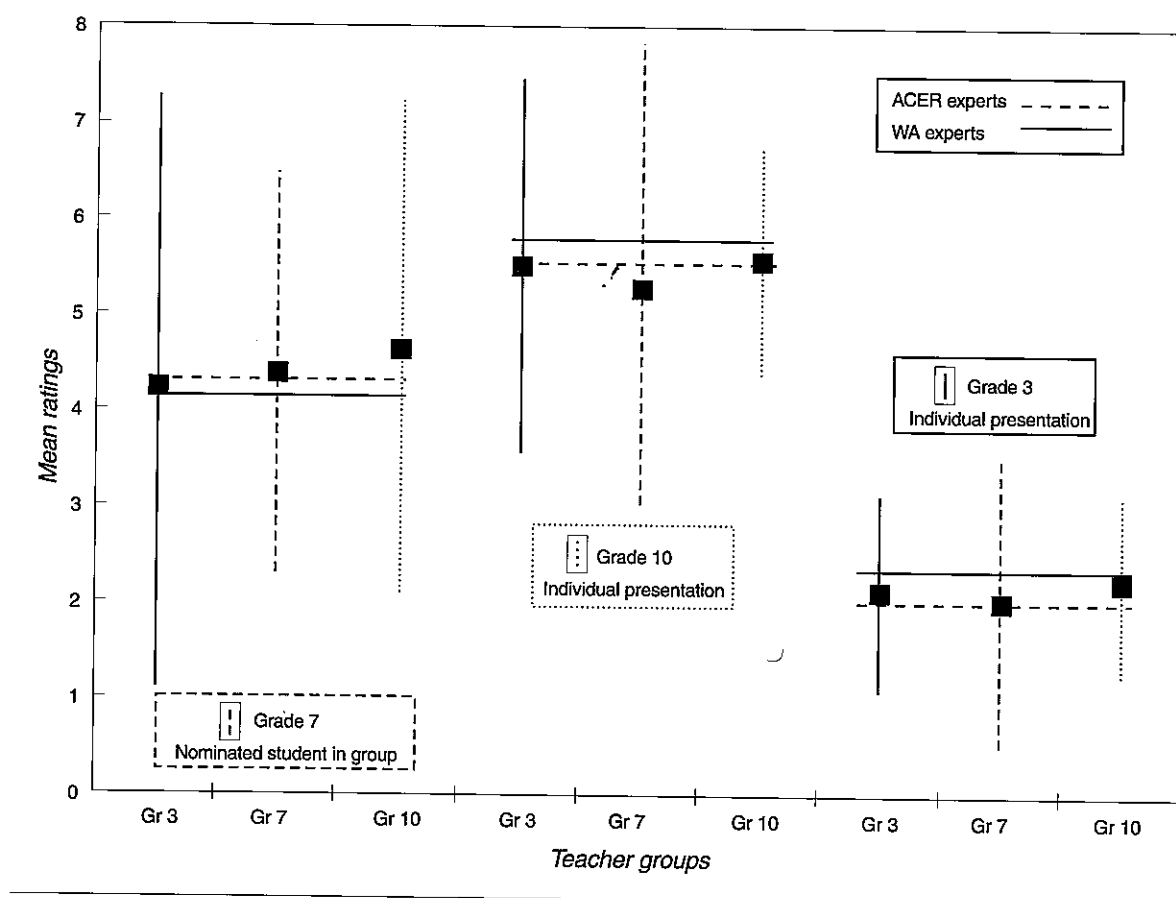


Figure 2. Mean ratings by teachers and 'experts' of speaking performances



- the performance of a nominated student in that Grade 7 group;
- an individual Grade 10 student presenting to the class; and
- an individual Grade 3 student presenting to the class.

The teachers were provided with a specific marking guide, using criteria for ideas or content, organization or focus, language usage, sense of audience and on-balance overall assessment. Each criterion was described at eight levels corresponding to the levels in the outcome sequences in the curriculum framework

There was considerable variation in the judgements of each student performance. The ranges of scores, on a 0-8 scale, are shown for each of the four performances in Table 3. Generally, the ratings of the teacher group most familiar with the grade level of the performer varied less than those of the other two teacher groups.

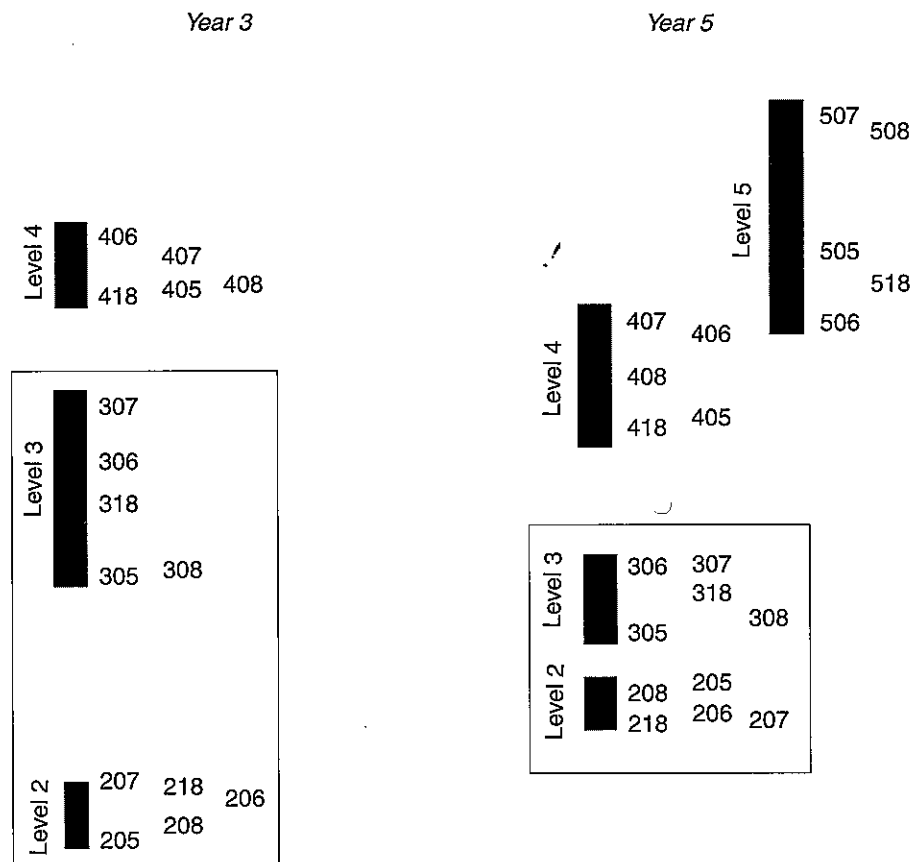
This is made clearer in the distributions of the ratings for the three individual performances, shown in Figure 1. The effect of less varied ratings is most marked with the Grade 7 teachers' ratings of the Grade 7 student and the Grade 10 teachers' ratings of the Grade 10 student.

Further examination of the data to investigate differences in rater severity (Mendelovits, 1997) revealed that the teachers were not consistent in their severity as judges. There was some weak evidence of teachers being more lenient at the grade level at which they teach, suggesting that they might set less realistic expectations for students at other levels.

Two expert groups also rated the sample student speaking performances. One consisted of the two test developers at ACER who had conceived and prepared the whole assessment package. The other was a group of six Western Australian teachers



Figure 3. Differences in judged difficulty of tasks by grade level of teacher



and researchers who had helped to develop or refine the student outcome statements for speaking and listening on which the marking guides were based. Some of them had been involved in the 1991 pilot project on oral language and some had participated as raters in the trial of the 1995 instruments. The ratings of these two groups are shown in relation to the ratings of the three teacher groups in Figure 2.

While the teacher ratings of these speaking performances are quite varied, their mean rating is close to the mean ratings of the experts. Further, the mean ratings for the teachers involved with the same grade as that of the student being judged are generally closest to those of the ACER expert judges.

In a case like this, where the results of individual students are used to monitor the system as a whole and not to report on individual students, the problem

of variation in teacher ratings is less serious than it would be in a high-stakes assessment situation where individual student results would matter. What the data do show, however, is that outcome statements in relation to performance are not unambiguous criteria for performance assessment in the hands of teachers.

There are other more general data on teacher assessment, drawn from trial work for the Australian National School English Literacy Survey (Management Committee for the National School English Literacy Survey, 1997; Masters and Forster, 1997), that point as well to inconsistency among teachers at different grade levels in their use of criteria for judging student work. The results in Figure 3 show the relative difficulty levels determined by Grade 3 and 5 teachers of tasks at different levels in the sequence of outcomes in the Australian national curriculum framework.



The Grade 3 teachers see a much greater difference between tasks at Levels 2 and 3, the ones with which they most frequently deal, than do the Grade 5 teachers. They also see the Level 4 tasks as more difficult, relative to the tasks at Levels 2 and 3, than do the Grade 5 teachers.

There is, however, some encouragement to be found in the full-national literacy survey that followed on from that trial work. The survey was also able to take account of the experience with the speaking performance assessments in the 1995 Western Australian study. With a more elaborate training programme for teachers, and a specialist group to train and assist them, a greater degree of consistency of judgements across schools was achieved (Management Committee ..., 1997).

Assessing experimental work in science

In 1998, the Western Australian Monitoring Standards in Education programme included science and again involved Grades 3, 7 and 10. All five strands of the science curriculum framework were assessed, viz.:

- Life and Living;
- Earth and Beyond;
- Energy and Change;
- Natural and Processed Materials; and
- Working Scientifically.

For the Working Scientifically strand, students were required to conduct simple experiments with materials provided, to make observations, to collect and organize data, and to interpret the data. There were link materials across all grade levels to permit the reporting of the results of all grades on a common scale. Preliminary analyses show that all of the responses fit a single scale satisfactorily but, because links are also being made with results from the 1993 monitoring of science, the two scales required for those data are being preserved for the 1998 analyses, viz.:

- Content scale
- Working Scientifically scale

Barry McGaw

The 1993 Working Scientifically scale was defined with open-ended, short-answer, written-response questions dealing with conceptual understanding of scientific concepts. The stimulus material for each question was a cartoon sequence requiring a written explanation (for example, of the movement of a skateboard). The 1998 Working Scientifically scale was defined by the tasks involving experiments (for example, measuring the distance an object is blown by air released from a balloon, mapping the range of peripheral vision, determining the extent to which different types of material can be stretched and reading aerial photographs of a mine site and township taken about a decade apart).

Assessing individual achievement

The English and science examples discussed above are from system-level monitoring during the compulsory years of schooling where individual students in a representative sample do not receive personal reports. In full cohort assessment programmes providing individual reports on students, the assessment programme is typically much more limited. Reporting on students' achievements in relation to the full range of curriculum outcomes is presumed to be the responsibility of the teachers, on the basis of their school-based assessments.

It is generally assumed that the relatively explicit specification of expected learning outcomes would enable teachers to provide essentially comparable judgements on the outcomes using only their local school-based assessments. The evidence from the investigations of performance assessments suggests clearly that such an assumption should be readily made. Careful specification must be accompanied by careful training it seems.

These issues come into sharpest focus in high-stakes assessments such as those at the end of secondary schooling, on which admission to university is based, and in various assessments for professional



accreditation. In many education systems this kind of high-stakes assessment is conducted by agencies external to the schools.

The use of external examinations imposes limitations on what can be assessed but the range of assessment practices can be made quite broad if the necessary resources are committed. A review of the New South Wales end-of-secondary Higher School Certificate documented some of this extension, introduced in an attempt to increase the curriculum fidelity of the external assessments (McGaw, 1996).

Some of these new forms of external assessment require response in a constrained setting at a fixed time in much the same way as examinations do, but not through written responses to examination questions. Other new forms permit sustained work on a product which is assessed when completed. Examples include the following.

- oral and aural tests:
 - listening in contemporary English;
 - speaking and listening in languages other than English; and
 - aural skills in music;
- a performance produced on a test occasion after sustained preparation of a known piece over a considerable period of time:
 - of individuals in music, dance, ballet; and
 - of groups in dance and drama;
- a final product produced through sustained individual work over a considerable period of time:
 - research projects in society and culture, agriculture, rural technology and food technology;
 - design projects in design and technology;
 - major works in visual arts;
 - compositions in music; and
 - extended essays in visual arts as one option in musicology in music.

These are all examples of external assessment in the same way that examinations are. The assessments are carried out by personnel external to the school and appointed

for the task, in this case, by the New South Wales Board of Studies. The assessments extend the scope of external examination in ways that address some of the disadvantages of traditional examinations.

Where assessment for these high-stakes purposes is supplemented further by the use of school-based assessments, various strategies are used to maximize comparability across schools in assessment practices and interpretations. One widely used strategy has been meetings of teachers at which samples of student work are exchanged and assessed to determine whether consistent judgements are being made across schools.

These practices have been most widely used at the end of compulsory schooling (for instance, the GCSE and its earlier equivalents in England and Wales, and various Grade 10 certificates in Australia). Where they have been used at the end of secondary level, there has been evidence of grade inflation as consensus among teachers in the meetings to achieve comparability of assessment had been obtained by lowering the standards required for particular assessments.

As a consequence, some systems have retreated to statistical controls to render assessments in different schools comparable. The most common strategy is to scale the results of school-based assessment (including performance assessment) to match the distribution of external examination results in that subject in the particular school. In the Australian State of Victoria, a different approach is taken. All students are required to take an externally set and marked General Achievement Test. The results from school-based assessments in each subject in each school are compared with those for the same students on the General Achievement Test. If the two sets of results are not comparable, personnel from the examination authority, who are external to the school, review the school-based assessment. The school-based results are not displaced by the results on the common external assessment, nor is their distribution



adjusted to that of the external assessment, but they are challenged and subjected to external review.

The desire and the challenge

Performance, in a variety of forms, is an important part of the outcomes desired in many curriculum areas to increase the validity of assessment. The challenge is to find ways to assess performance that give it the appropriate priority, whether in programmes to monitor the performance of education systems or in high-stakes assessment of individual students. This validity can only be achieved if the assessment tasks match the requirements of the curriculum and if the assessment results have a sufficient level of reliability.

Careful specification of the kinds of performances to be assessed and the criteria to be used in judging the performances will not, of themselves, produce reliable assessments. There is more to be learned about how best to achieve this but, so far, it is clear that careful training of the judges, whether they are classroom teachers or external examiners, is essential.

References

- COOK, J.; RANDALL, K.; RICHARDS, L. 1997. *Student Achievement in English in Western Australian Government Schools: 1995*. Perth, Education Department of Western Australia.
- MANAGEMENT COMMITTEE FOR THE NATIONAL SCHOOL ENGLISH LITERACY SURVEY. 1997. *Mapping Literacy Achievement: Results of the 1996 National School English Literacy Survey*. Canberra, Department of Employment, Education, Training and Youth Affairs.
- MASTERS, G. N.; FORSTER, M. 1997. *Literacy Standards in Australia*. Canberra, Commonwealth of Australia.
- MCGAW, B. 1996. *Their Future: Options for Reform of the Higher School Certificate*. Sydney, Australia, Department of Training and Education Co-ordination.
- MENDELOVITS, J. 1997. *Evaluating the Reliability of Teachers as Raters in A Performance Assessment Programme*. (Paper presented at the Annual Conference of the American Educational Research Association, Chicago, Ill.)

Comments on 'Performance assessment'

Samuel J. Messick

Introduction

McGaw began his paper with an important question. What constitutes performance assessment?

One view is that there is a continuum of tasks that go from simple-answer, open- or constructed-response formats to

more complex demonstrations. What constitutes the stimulus for the task is a great issue. Although a constructed-response item, unlike a multiple-choice item, does not have five options from which to select, it can have a single stem.

Constructed-response items become increasingly open as we go from simple answers to demonstrations, port-

Samuel J. Messick

folios, exhibits and so forth, at the other end of the continuum. Some may say that a constructed-response test can be considered anything beyond multiple-choice. As McGaw notes, that is a very broad definition.

■
■ **Issues in performance
assessment**
■
□

Performance assessments, the way I see them, are subsets of constructed-responses that are extended performances that lead to a product or that track a performance. Simulation assessments are a subset of performance assessments that are either actual criterion samples or close simulations of criterion requirements in terms of the nature of the task and the resources (including time and materials) required to approximate a real-world performance. It is not necessarily that they are real-world criterion samples, but they are close simulations. Even if they were criterion samples, they are not the same as criterion performance because there are other issues now involved, such as how anxiety might affect performance in ways that would not occur in the real world.

Washback

We must also consider the issue of 'washback'. Using a test that assesses test-takers' ability to speak in a foreign language as an example, the way you test has a 'washback' effect on the way learning occurred. It is somewhat like the Frederiksen and Collins notion of systemic relationship: the nature of the test itself brings out positive facts about the way teaching and learning occurs. In this notion of positive 'washback', we would want whatever test-takers do to prepare for the test to be transparent with respect to whatever they do to learn language. If the test requires test-takers to do things that are not required in learning language, that is probably a source of irrelevant difficulty. So we would like the lan-

guage to be as authentic as possible in that sense, so that there is no distinction between the test task and the learning task.

Generalizability

Another serious issue is generalizability. If you are going to evaluate a performance in a particular way, using the Olympic Games as an example, the issue is not to generalize from a single performance and claim that a contestant can dive that well or swim that well or dance that well tomorrow, or the next day, or a month from now. Generalizability is not an issue because it is the 'performance-as-target' that is being evaluated. In education we rarely use performance assessments as targets. We use them as vehicles for the assessment of knowledge and skills and their attributes. Here, generalizability is an issue because we want to make generalizable statements about the nature of the skill level, work, or other aspect of the performance of the student. And that's an important distinction, because generalizability is the soft underbelly of performance assessment. Being able to assess ability and get sufficiently generalizable results is one of the difficulties that we face. It is difficult to get an assessment where you can afford to have many extended tasks, so you are limited to one or two samples. That's usually the case in large-scale assessments. However, there are some special studies that look at the issues related to generalizability.

McGaw cited a study of the scoring assessment of an oral language in Western Australia. The study showed considerable variation in the scores given by raters. In an ETS study on writing with a different focus, participants were each asked to write six essays; two essays on different topics in each of three different modes: narrative, expository and persuasive. The question was: Were the scores generalizable across modes and topics? It was found that the mode made a difference in the score and the topic did as well. Although the combined score of the six essays was reliable,

the scores on the individual essays varied. In brief, it was not possible to generalize from one sample to another. This tells us that trying to generalize from a test score on a single 20-minute writing task is very problematic. This does not mean that we can't test writing. It does mean that we need to ensure that we get an adequate sample. This makes generalizability an issue when using performance assessments in large-scale assessments.

Approach

The approach used is another issue. Is the assessment the target of what is to be assessed or is it the vehicle for assessment? It is the notion of whether you take a task-centred approach or a construct-centred approach. The task-centred approach has some difficulties. It is based on the premise that some tasks are worth learning to perform and some are not. Does this mean that there is a set of working tasks that sample ability and that can produce scores from which we can generalize?

In a task-centred assessment, if the task seems to be important, people will think it is an excellent test. On reflection, they may ask what constructs underlie the task. That's not a very good way to proceed.

I would argue that it is better to start with the constructs and that the constructs should come from the content standards. Start with the construct and then ask what kinds of tasks will reveal that construct at sufficient levels of complexity. The construct-oriented approach not only helps us to select and develop tasks in a rational way, it also helps to develop scoring rubrics.

Appropriate use

Performance assessment is certainly not new. It is what we used to do before we had

multiple-choice tests. During the Second World War, Harold Gulliksen and Norman Frederiksen were asked to see if they could improve some of the selection procedures that were used to assign Navy recruits to training programmes. In the beginning of the war we weren't doing so well and it was important to get people trained quickly for important military assignments. The selection of the right people for the right tasks was an important issue. Gulliksen and Frederiksen looked at the gunner's mate test. In reviewing the selection procedures, they found that the best predictor of performance in the training programme was reading comprehension. This puzzled them, until they looked at the nature of the criterion examination and found that it was multiple-choice and based on the manuals for various guns. They revised the examination and made it a hands-on performance examination. The examinees had to disassemble guns and then reassemble them. Mechanical aptitude proved to be much better assessment than reading comprehension for predicting individuals' aptitude for performing the duties of a gunner's mate.

As McGaw suggests, performance-based assessment is not appropriate in all circumstances. However, it is clear that it has some important uses. In the gunner's mates case, it is credited with saving lives.

Bibliography

-
-
-
- FREDERIKSEN, J. R.; COLLINS, A. 1989. A Systems Approach to Educational Testing. *Educational Researcher*, Vol. 18, No. 9, pp. 27-32.
- FREDERIKSEN, N. (with Lt. A. E. Monroe). 1945. *The Development of Achievement Tests for Gunner's Mates Schools*. Washington, D.C., United States Department of Commerce.

Comments on 'Performance assessment'

Robert Linn

Defining performance assessment

Barry McGaw distinguished several uses of the term 'performance assessment'. The broad definition that has frequently been used in the United States would include any constructed-response assessment, no matter how simple the nature of the response. This sweeping view of performance assessment comes about, as McGaw implies, because of the widespread use of the multiple-choice format in the United States. Indeed, performance assessment is often the catch-all for anything that is not multiple-choice or a related fixed-response, machine-scorable format. I agree with McGaw that this meaning of performance assessment is too broad to be very useful.

While rejecting the broad concept of performance assessment as anything other than multiple choice, McGaw also rejects as too narrow the very restrictive definition of performance assessment as applying to only those performances that occur on the job or in real-world settings outside of school. Again, I concur.

Trimming the extreme interpretations helps, but still leaves us with a somewhat fuzzy concept. I would include written essays and other types of written responses that are sometimes referred to as extended-response items as examples of performance assessments. As McGaw's examples of assessments of student speaking skills illustrate, performance assessments need not be paper-and-pencil tests. In the area of science, performance assess-

ments involving more than paper-and-pencil exercises are sometimes signaled by the qualification of 'hands-on'. Hands-on performance assessments in science generally involve some manipulation of materials and instruments, measurement, observation or recording. In addition, they may well involve written statements of hypotheses, results and conclusions, but the overall assessment involves more than paper-and-pencil exercises.

The nature of the performance to be assessed should, as McGaw suggests, be determined by the construct that the assessment is designed to measure. This is evident in McGaw's illustration of speaking as the target of the assessment. As he states, 'with speaking there could be no substitute for performance assessment'. There simply are no plausible surrogates for speaking.

In some areas surrogate measures may work quite well in place of more costly and time-consuming performance assessments. Caution is needed, however. There is a good deal of evidence which suggest that method variance can distort the validity of inferences about the construct of interest. Rich Shavelson and his colleagues (for example, Shavelson et al., 1992) have shown, for example, that even apparently high fidelity computer simulations of science tasks may be only weakly related to actual hands-on performance assessments involving the same conceptual task. Measured proficiency in completing an electrical circuit in a computer simulation, for example, may not be an adequate substitute for measures of proficiency using actual batteries, wires and light bulbs.

Scoring

An obvious characteristic of the types of performance assessments that McGaw discussed is that the performances must be scored by human judges. Judges clearly are a source of measurement error in performance assessment. The magnitude of the error either in the form of a main effect for judge leniency or in the form of an interaction with the examinee depends on many factors, including the nature of the performances being scored, the specificity of the scoring rubrics and the training of judges to use common criteria. Early in this century Starch and Elliott attacked constructed-response tests by reporting the extraordinarily wide variation in the marks teachers of English and teachers of mathematics assigned to the same student papers (Starch and Elliott, 1912; 1913). Their results were used to garner support for the then 'new' objective tests. Those results, however, were an unfair test of the potential value of judgmentally scored performance assessments. The Starch and Elliott studies had two important flaws: clearly defined scoring rubrics were not used to score the assessment products; and the teachers used to score the assessments were not trained in applying common criteria when judging the student work.

The teacher ratings of student speaking reported by McGaw had the benefit of a 'specific marking guide' with eight levels. Nevertheless, the spread of scores assigned to a given student when different teachers scored speaking performance was still quite wide. It is encouraging that teachers who teach at the grade of the student whose performance is being scored show a smaller spread of scores than teachers teaching at other grade levels. However, the spread of scores for teachers most familiar with the student's grade level was still substantial. What constitutes an acceptable level of agreement among judges depends heavily on the uses to be made of the results. For example, a level that may

be adequate for purposes of reporting aggregate results or for low-stakes uses in the classroom may not be adequate if the scores are to be used to make decisions of real consequence (e.g. retention in grade or assignment to a remedial programme) about individual students.

Additional training in scoring or the use of moderation procedures might be expected to improve the level of reliability. However, if the results are to be used for high-stakes accountability purposes, it might be necessary to use multiple ratings or audits, with the possibility of rescoring.

Simply abandoning performance assessment and retreating to the safer haven of multiple-choice tests is not the solution to the problem of scoring reliability. In doing so, the gain in reliability may be more than offset by loss in validity.

Validity

The first two questions for a performance assessment are the same questions that should be asked of any assessment. What is the purpose of the assessment? What is the construct that needs to be measured? Under the first question there are many subquestions, such as: How will the results be reported and used? What decisions will be made about individual students based on the assessment results? Who is expected to make what uses of the results?

In education assessments the specification of the construct usually starts with the identification of the curriculum framework or content standards that define what it is students are expected to learn and to be able to do. Too often, the specification of the construct ends with the identification of the curriculum or content standards. This is unfortunate because curriculum frameworks and content standards are usually too general to clearly define the construct that is the intended focus of the assessment. Elaborations of content standards and curriculum frameworks are needed to guide the choice of assess-

ment tasks and the range of performances that need to be judged. The elaborations need to go beyond statements of broad goals and objectives to the specification of the types of performances desired as the result of instruction.

■ ■ ■ Constraints and uses ■ ■ ■

It is, of course, easy to say that the tasks as well as the nature of the performances that are most relevant for an assessment should be determined first by determining the purpose of the assessment and the construct to be measured. In practice, however, there are multiple constraints to be considered. Cost, both in monetary terms and in terms of time required of students and teachers, can be a major determinant of what can realistically be assessed. In large-scale, external assessment programmes, cost may sharply limit what can be assessed.

More ambitious assessments involving extended performances are most apt to be feasible under one of the following circumstances. First, the assessment is an integral part of instruction and learning and is under the control of teachers. Second, the performances to be assessed are so critically important that the cost of the assessment can be readily justified. Licen-

sure and certification tests for physicians and airplane pilots are obvious examples of this second category. Examples in this category for public school education are harder to come by. Third, assessment goals of monitoring system performance or school accountability can be achieved by using matrix-sampling procedures.

McGaw has raised some important issues regarding the development and evaluation of performance assessments. As he suggested, the challenge is to find ways to give appropriate priority to performance assessments within the constraints of the purposes to be served by the assessment and a clear understanding of the constructs to be assessed.

■ ■ ■ References ■ ■ ■

- SHAVELSON, R. J.; BAXTER, G. P.; PINE, J. 1992. Performance Assessments: Political Rhetoric and Measurement Reality. *Educational Researcher*, Vol. 21, No. 4, pp. 22-7.
- STARCH, D.; ELLIOTT, E. C. 1912. Reliability of Grading High School Work in English. *School Review*, Vol. 20, pp. 442-57.
- . 1913. Reliability of Grading High School Work in Mathematics. *School Review*, Vol. 21, pp. 254-9.

Comments on 'Performance assessment'

Mark D. Reckase

■ ■ ■ The Information Bottleneck: how much information can we obtain from a student in a fixed period of time? ■ ■ ■

The presentation by McGaw provides a very good introduction to the topic and also pro-

vides some tantalizing data about the process of evaluating the results of performance assessments. While listening to the McGaw presentation, I was reminded of some early work by one of the pioneer researchers in the telephone industry. This researcher, Claude Shannon (1949), deter-

mined that there is a physical limitation to the amount of information that could be sent over a phone line. Once this physical limitation has been reached for a particular wire, the only thing that can be done to increase the amount of information going down the line is to use a bigger wire, use more wires, or change the transmission technology. The use of fiber-optic cable is an example of using a bigger wire, and the use of laser light for the transmission of information is an example of using new transmission technology.

The reason for presenting this seemingly tangential information is to suggest that we have the same information transmission challenge in the education assessment arena as is encountered by the telephone industry. We are trying to get students to transmit information to us about the knowledge and skills that they have gained through the education system. To get this information, we have to use a communication channel. In this case, the communications channel is the means that the students use to provide us with the information we need. The channel could be the process of making dots on an answer document. It could be writing an essay on the topic. It could be doing a science experiment that is videotaped for later evaluation.

Some important questions about the channel are: How much information can we get through that information channel per unit of time? How can we get the information we want, given the physical limitations in the communications channel? For each type of physical act used by the examinee to communicate knowledge and skills, there is a physical limitation for the amount of information that can be provided by the student per unit of time.

During the one-hour Mathematics Test on the ACT Assessment college entrance examination (ACT, Inc., 1997), each student provides 60 bubbled responses. If the information about the particular response selected from the multiple choices is not used, but only the correctness or incorrectness of the response is con-

sidered, then the 60 responses result in 60 binary digits. The fact that they are binary digits is not a problem. Very detailed information can be transmitted using binary digits. For example, space probes have transmitted pictures of Jupiter to Earth using binary digits. The issue is more about how many binary digits are needed to give a reasonable representation of the knowledge and skills that have been acquired by the student. The production of these binary digits is one communication channel that is used to gain information about student capabilities. Does it provide the information we need to know, and is it an efficient way to get that information?

An alternative communication channel to producing binary digits is producing and scoring performance assessments. In the performance assessments described by McGaw, the scored result for an hour-long activity could be a single performance rating on a 0-to-8 rating scale. This would appear to provide much less information than that provided by 60 binary digits. Of course, the complex work that is performed by the student could be evaluated using multiple rating scales so that more information is acquired during the period of time. A question for measurement theorists is whether responding to multiple-choice questions or to performance exercises, with subsequent ratings of the quality, can provide more information about student capabilities. McGaw clearly indicates he believes that performance assessments provide less information than multiple-choice tests provide, at least in terms of content coverage, when used to assess every student in a population. He indicates that the monitoring is 'impoverished' in comparison with a thorough evaluation of achievement in a curriculum area.

The lack of efficiency in the coverage of curriculum may not be an issue, depending on how the performance assessment results are to be used. If an in-depth analysis of students' skills is desired, but every student does not have to be evaluated on every concept, McGaw argues that

matrix sampling procedures can be used to control the amount of testing time per student. When matrix sampling is used, each student interacts with a relatively small number of performance assessment tasks and the results are aggregated over all students to provide group reporting of results. Also, if the skills and knowledge that are measured by the performance assessment are different than those measured by the multiple-choice test and those different skills are highly valued, then the inefficiencies in the performance assessment approach are of less importance.

■
 ■ **Performance assessment
 as an instructional goal**
 ■
 □

There is another way that performance assessments may be valued beyond the value of the measurement information they provide. A number of educators have argued that multiple-choice tests are not good models for the types of behaviour that we want students to exhibit (Mitchell, 1992; National Commission on Testing and Public Policy, 1990). Multiple-choice tests are composed of short tasks that have correct and incorrect answers. Rather than this type of artificial task, these educators are looking for more realistic tasks that are similar to the type of work students are expected to do as an outcome of the education process. These tasks have been labeled as 'authentic', as McGaw indicates. Rather than provide indicators of performance, performance assessments are expected to provide good models for performance and to give students the opportunity to demonstrate their ability to match these good models.

■
 ■ **Definition of
 performance assessment**
 ■
 □

McGaw takes the middle ground on the question of what constitutes performance assessment. He favours student work that is

extended and that provides good models for instruction, but he does not emphasize that approach so much that only real-world activities are included. My own attitude toward performance assessment is very similar to McGaw's. Performance assessments are both more than open-ended items and less than real-world tasks. The characteristics of performance assessment exercises are shown in a practical way by the PASSPORT portfolio assessment system (Reckase, 1995). The performance assessment tasks used by this system are actual class work assignments, but the tasks are selected to match rigorous requirements related to curriculum standards and as good models for instruction. The requirements for the tasks are listed in 'work sample descriptions' that provide a loose framework for the selection of assignments, but also limit the tasks to those that can be rigorously scored. I believe this balance between flexibility and rigour, and focus versus breadth, meets the conditions that McGaw has specified. The key is that there is enough structure to the requirements for the tasks that they can be properly scored. Given McGaw's concerns about scoring performance assessments, this perspective is likely to be consistent with his.

The portfolio approach to performance assessment also shows promise for increasing the information provided about student capabilities by changing the capacity of the information channel. Rather than collecting information from relatively short, timed exercises, work can be collected from an entire school year. This approach has the added advantage of considering a broad spectrum of student work rather than a restricted sample.

■
 ■ **Scoring performance
 assessments**
 ■
 □

Performance assessments are different from multiple-choice assessments in that the scoring of these assessments is often as challenging as producing the work that is being

scored. Persons scoring the performance assessment documentation must be expert enough to be able to carefully evaluate the work, but they must also be willing to score the materials according to the detailed rules that have been provided by the designer of the performance assessment tasks. McGaw's paper shows the criticality of the training and the influence of the scorers' background on their performance as scorers. The variance of ratings for teachers who were rating students at the grade level that they usually taught was less than the variance of ratings for levels that they did not usually teach. This was the case even though the teachers were all given the same training. He also showed that teacher perceptions of task difficulty varied with the grade level that the teachers typically taught. These are important findings. They imply that the knowledge that teachers have about the capabilities of the students has an effect on the scoring process.

One interpretation of this result can be constructed from a true score theory conceptualization. Assuming the true variance of student performance is constant, the fact that the teachers who are less familiar with the level of student tend to have higher variance of scores than those who regularly teach students at those levels suggests that the less familiar teachers have greater error in their scoring than the more familiar teachers. This further implies that the scoring reliability for teachers who are evaluating students who are not at the grade level that they teach is less than that for teachers who teach at that level. The policy implication is that the selection of appropriately qualified scorers is an important part of the scoring process. Good training is not enough. Scorers must have experience that is appropriate for the task and for the student population.

Summary

McGaw has clarified the definition of performance assessment and has identified a

critical component of this type of assessment procedure: scoring. The research he reports provides compelling evidence that rater characteristics must be considered when devising the scoring process for performance assessments. The concept of information channel was presented to provide a framework for understanding the challenges to the use of performance assessment. Those challenges relate to the efficiency of acquiring information about students' capabilities. Since performance assessments are relatively inefficient compared with other alternatives, their use must be supported on grounds other than the amount of information they provide. Possible justifications for the use of performance assessments are that they provide models for good work and they assess unique skills that are not tapped by other assessment procedures.

References


- ACT, Inc. 1997. *ACT Assessment Technical Manual*. Iowa City, Iowa.
- MITCHELL, R. 1992. *Testing for Learning: How New Approaches to Evaluation Can Improve American Schools*. New York, The Free Press.
- NATIONAL COMMISSION ON TESTING AND PUBLIC POLICY. 1990. *From Gatekeeper to Gateway: Transforming Testing in America*. Chestnut Hill, Mass., Boston College.
- RECKASE, M. D. 1995. Practical Experience in Implementing a National Portfolio Model at the High School Level. *NASSP Bulletin*, Vol. 79, No. 573, pp. 31-6.
- SHANNON, C. E. 1949. The Mathematical Theory of Communication. In: C. E. Shannon and W. Weaver (eds.), *The Mathematical Theory of Communication*. Urbana, Ill., University of Illinois Press.

Mark D. Reckase



5. Purposes and challenges of international comparative assessments

Tjeerd Plomp



Introduction

This paper aims to consider and illustrate some of the issues related to the purposes and uses of international assessment, with particular reference to the type of studies conducted by the International Association for the Evaluation of Educational Achievement (IEA).

IEA is the organization that conducts international comparative studies in which education achievement is assessed in the context of process and input variables. IEA's mission is to contribute, through its studies, to enhancing the quality of education.

Over almost forty years, IEA has developed as a co-operative of research institutes representing, at present, fifty-five education systems. Many countries are now represented in the IEA General Assembly by policy-makers. National research co-ordinators and centres for IEA studies are often the most prominent ones in their country; some are part of their respective ministries of education, while others are linked to universities or are independent research centres. By its nature, IEA provides a network of institutes and individuals, which together represent a great deal of experience and

intellectual capacity. As such, it is a meeting place for policy-makers, educators, scientists and researchers.

Over the years, IEA has conducted many survey studies of basic school subjects. Most of them were curriculum-driven. That is, a test grid for measuring education outcomes was developed based on an analysis of the curriculum of the participating countries. All these studies also included instruments to measure school and classroom process variables, as well as teacher and student background variables. Some examples of the studies conducted are mathematics and science, reading literacy, civics education, and English and French as foreign languages.

IEA also conducts studies that are not curriculum-based. Examples are the Pre-Primary Project and the Computers in Education Study, of which a successor (the Second Information Technology in Education Study, SITES) is under preparation.

TIMSS is the largest international comparative study of education achievement ever undertaken. The TIMSS achievement testing in mathematics and science included:

- 45 countries;
- five grade levels (3rd, 4th, 7th, 8th and final year of secondary school);
- more than half a million students;

- testing in more than 30 different languages;
- more than 15,000 participating schools;
- nearly 1,000 open-ended questions, generating millions of student responses;
- performance assessment;
- questionnaires for students, teachers, and school principals containing altogether about 1,500 questions; and
- many thousands of individuals to administer the tests and process the data.

TIMSS was conducted with attention to quality at every step of the way. Rigorous procedures were applied to translate the tests and numerous regional training sessions were held in data collection and scoring procedures. Quality control observers monitored testing sessions. The samples of students selected for testing were scrutinized according to rigorous standards designed to prevent bias and ensure comparability. Countries that did not meet all the quality criteria were marked as such in the published tables of the TIMSS results.

The achievement results of TIMSS have been published by the International Study Centre at Boston College, United States of America. (See the references for a full list of publications from this study.) Some of these results are summarized and discussed here to illustrate the potential richness of international comparative assessment studies.

IEA is repeating TIMSS for Grade 8 in 1998 for the Southern Hemisphere and in 1999 for the Northern Hemisphere. A number of countries that did not previously participate in TIMSS will be able to join the TIMSS-repeat study thanks to World Bank support.

The second study IEA is undertaking is the Civics Education Study (CES). This study finished its first phase, the development of country profiles, in 1998, and is collecting data at the school, teacher and student levels in early 1999.

Another ongoing study, different in scope, is the Pre-Primary Project, a study of policies and practices in early childhood care and education.

Tjeerd Plomp

A second information technology in education study (SITES) was started in the fall of 1997 with an indicators module. Two other modules are also being planned, namely a module of international comparative case studies of innovative practices in the use of information and communication technology, and a survey of schools, teachers and students in 2001.

Although many people accept the merit of studies that focus on purely national or regional concerns, they question the benefits of carrying out comparative research. They question whether it is possible to make cross-national comparisons that are sensitive to, and reflect, differences between the curricula, structures and stages of education development of the participating countries.

Arguments put forward by Foshay (1962) and others suggest that while comparisons between education systems must indeed be viewed against a complex background, they do have the advantage of highlighting the similarity of education problems and issues between and among nations. Foshay observes that

if custom and law define what is educationally allowable within a nation, the educational systems beyond one's national boundaries suggest what is educationally allowable (p. 2).

Comparative perspectives also allow us to examine the impact and effect on education systems of policies that are applied consistently within nations but may vary across nations. The understandings we obtain from cross-national comparisons of policies such as age of school entry, hours and methods of instruction, and teacher training can provide us with new insights into the performance of our own education system in general, and of the relationship between student performance and its antecedents and consequences in particular.

IEA recognizes two purposes of international comparative achievement studies:

- to provide policy-makers and education

practitioners with information about the quality of their education systems in relation to relevant reference groups; and to assist in understanding the reasons for observed differences between education systems (which serves policy-makers' needs, but is clearly of interest to researchers).

In line with these two purposes, IEA strives in its studies for two kinds of comparisons.

The first consists of straight international comparisons of effects of education in terms of scores (or subscores) on international tests, as illustrated for TIMSS in Table 1 and Figure 1.

The second kind of comparison concerns how well a country's intended curriculum (what should be taught in a particular grade) is implemented in the schools and achieved by students. This kind of comparison focuses mainly on national analyses of a country's results in an international comparative context.

IEA was founded as a research co-operative with primarily an academic research focus. Since the beginning of the 1980s, however, IEA has begun to focus more closely on the interests of policy-makers. The fifth edition of OECD's *Education at a Glance* (1997) presents a number of indicators based on the TIMSS results. Examples of IEA publications that address relevant policy questions are Postlethwaite and Ross (1994) and Keeves (1996); another relevant source is Kellaghan (1996).

Although not every study should have a size and a design as comprehensive as TIMSS, IEA believes that the conceptualization and design of its studies allows for analyses that meet the needs of both policy-makers and education practitioners.

Functions of IEA studies

The relevance of IEA studies reaches much further than just making straight comparisons in the form of league tables. The fol-

lowing functions illustrate the importance of international comparative achievement studies (and of education indicators).

Description: mirror function

To provide policy-makers and the education community with information about the status of their education system in an international comparative context: this function is considered by many to be interesting in itself. Many policy-makers have now recognized that such information is a good starting point for generating questions for in-depth analysis. This can be illustrated with some exemplary results from TIMSS presented in Table 1 and Figure 1 (also discussed in Plomp, 1997). Table 1 contains achievement test results for science in Grades 7 and 8, while Figure 1 presents the multiple comparisons for Grade 8 mathematics achievement.

Table 1 and Figure 1 illustrate one of the purposes of international comparative achievement studies, namely providing policy-makers and education practitioners with information (indicators) about the quality of their education system in relation to relevant reference groups of similar nations. This is the 'mirror' function. Countries can determine whether or not they like the picture or profile of their country as compared to other countries.

Table 1 just gives 'horse race' data, with, for example, England with Grade 8 science in 10th place and Grade 7 science in 11th place. Figure 1 provides more information, namely a country can see which countries have mean achievement scores that are significantly lower or higher than their own, or that do not have statistically significantly different scores.

This type of information informs policy-makers in England how well their country is doing in comparison with other countries. It shows also that league tables like Table 1 contain limited information and may result in misleading interpretations, as they do not reflect any statis-

Table 1: Distribution of achievement in the sciences for students in the final years of middle school

| <i>Eighth grade¹</i> | | <i>Seventh grade¹</i> | |
|---------------------------------|--------------------------------|----------------------------------|--------------------------------|
| <i>Country/ territory</i> | <i>Average achievement</i> | <i>Country/ territory</i> | <i>Average achievement</i> |
| Singapore | 607 | Singapore | 545 |
| Czech Republic | 574 | Republic of Korea | 535 |
| Japan | 571 | Czech Republic | 533 |
| Republic of Korea | 565 | Japan | 531 |
| <i>Bulgaria</i> | 565 | <i>Bulgaria</i> | 531 |
| <i>Netherlands</i> | 560 | <i>Slovenia</i> | 530 |
| <i>Slovenia</i> | 560 | Belgium (Fl) ² | 529 |
| <i>Austria</i> | 558 | <i>Austria</i> | 519 |
| Hungary | 554 | Hungary | 518 |
| England | 552 | <i>Netherlands</i> | 517 |
| Belgium (Fl) ² | 550 | England | 512 |
| <i>Australia</i> | 545 | Slovakia | 510 |
| Slovakia | 544 | United States | 508 |
| Russian Federation | 538 | <i>Australia</i> | 504 |
| Ireland | 538 | <i>Germany</i> | 499 |
| Sweden | 535 | Canada | 499 |
| United States | 534 | Hong Kong | 495 |
| <i>Germany</i> | 531 | Ireland | 495 |
| Canada | 531 | <i>Thailand</i> | 493 |
| Norway | 527 | Sweden | 488 |
| New Zealand | 525 | Russian Federation | 484 |
| <i>Thailand</i> | 525 | Switzerland | 484 |
| <i>Israel</i> | 524 | Norway | 483 |
| Hong Kong | 522 | New Zealand | 481 |
| Switzerland | 522 | Spain | 477 |
| <i>Scotland</i> | 517 | Scotland | 468 |
| Spain | 517 | Iceland | 462 |
| France | 498 | <i>Romania</i> | 452 |
| <i>Greece</i> | 497 | France | 451 |
| Iceland | 494 | <i>Greece</i> | 449 |
| <i>Romania</i> | 486 | Belgium (Fr) ⁴ | 442 |
| Latvia ³ | 485 | <i>Denmark</i> | 439 |
| Portugal | 480 | Iran, Islamic Rep. | 436 |
| <i>Denmark</i> | 478 | Latvia ³ | 435 |
| Lithuania | 476 | Portugal | 428 |
| <i>Belgium (Fr)⁴</i> | 471 | Cyprus | 420 |
| Iran, Islamic Rep. of | 470 | Lithuania | 403 |
| Cyprus | 463 | <i>Colombia</i> | 387 |
| <i>Kuwait</i> | 430 | <i>South Africa</i> | 317 |
| <i>Colombia</i> | 411 | | |
| <i>South Africa</i> | 326 | | |

1. Eighth and seventh grades in most countries.
2. Flemish-speaking schools only.
3. Latvian-speaking schools only.
4. French-speaking schools only.

Countries shown in italics did not satisfy one or more guidelines for sample participation rates, age/grade specifications, or classroom sampling procedures.

Source: adapted from A. E. Beaton et al., *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS)*, Tables 1.1 and 1.2. Chestnut Hill, Mass., Center for the Study of Testing, Evaluation and Educational Policy, Boston College, 1996.

Table 2: TIMSS – Mathematics: England versus other countries

Significantly higher achievement:

| | | |
|--------------------|-------------|--------------|
| Singapore | Switzerland | Russian Fed. |
| Korea, Republic of | Netherlands | Australia |
| Japan | Slovenia | Ireland |
| Hong Kong, China | Austria | Canada |
| Belgium (Fl) | France | Belgium (Fr) |
| Czech Rep. | Hungary | Sweden |
| Slovakia | | |

No significant difference:

| | | |
|----------|-------------|---------------|
| Thailand | New Zealand | United States |
| Israel | Norway | Scotland |
| Germany | Denmark | Latvia (LSS) |

Significantly lower achievement:

| | | |
|---------|-----------|----------|
| Spain | Romania | Cyprus |
| Iceland | Lithuania | Portugal |
| Greece | | |

tical information. Figure 1, which does reflect this type of information, shows not only that England is not really performing better than the United States or worse than Germany, but also that European Union (EU) partners such as Ireland, Belgium, the Netherlands, France and so forth are performing significantly better. However, information from tables and figures like those presented here does not help policy-makers, curriculum developers and education practitioners understand why their education system is performing as it does; for example, why is England performing more poorly than many of its EU partners? The broad interest worldwide in the TIMSS results illustrates the relevance of this function.

Benchmarking

This function can best be illustrated with an example. Within TIMSS, some Asian countries, as well as Flemish Belgium and the Czech Republic, have the highest test scores in mathematics. If another country

is interested in improving its education in mathematics, it can compare its own situation with that of the Asian countries and/or the above-mentioned European countries, using any of the many variables related to curricular aspects of mathematics and science education, including curricular materials, pedagogical approaches and instructional processes, school variables, teacher background, teacher training and in-service training. Such analyses may result in proposals for change, although no easy answers can be expected. For such countries, an important question in a subsequent IEA study would be whether it is then performing closer to the reference countries chosen.

Monitoring of quality of education

One step further than benchmarking is monitoring: the regular assessment of education processes on different levels in the education system with the purpose of bring-

ing about change when and where needed ('informed decision-making'). This function is an example of assessment-led monitoring of the curriculum (but in the case of IEA studies, on the basis of curriculum-based assessment). For this use, trend data are needed, that is a cycle of regular assessments in the subject areas being monitored (like the IEA and OECD cycle of studies in mathematics, sciences and reading literacy). It is for this reason that IEA was asked to repeat TIMSS for the Grade 8 population in 1999.

Understanding observed differences

Policy-makers may want to understand differences between or within education systems from the perspective of national policy-making (this function should be distinguished from the next one: cross-national research).

This function is again one step beyond collecting data for monitoring purposes. It ultimately serves policy-makers' needs, but is clearly also of interest to researchers. To realize this function, information about learning and teaching processes and their inputs is required, along with in-depth analysis of achievement results in the context of these background data. IEA studies do collect different kinds of background data as well, but IEA considers this type of analysis an important task of the participating countries themselves, as they can determine which research and analysis questions are most relevant for their respective education systems. A good example is the analysis of the United States data in the IEA Second International Mathematics Study (SIMS), which resulted in a monograph entitled *The Underachieving Curriculum* (McKnight et al., 1989). However, no easy answers can be expected on what measures should be taken to improve education in a country. But this kind of research can lead to policy decisions about changes in education ('informed decision-making'), or to initiatives such as those in the United

States of America, where the National Council for the Teaching of Mathematics developed the well-known standards for the teaching of mathematics.

Cross-national research

This function refers to exploratory and/or in-depth research of the IEA databases. In TIMSS, this in-depth analysis is still to be done. Many examples can be found in the IEA volumes. Here we mention just two other examples. Postlethwaite and Ross (1994) did an exploratory research of the IEA Reading Literacy database (data collection in 1990-91) in an effort to find indicators discriminating between more effective and less effective schools in reading. The second example is Keeves' (1996) monograph *The World of School Learning: Selected Key Findings from 35 Years of IEA Research*, in which he discusses ten key findings with suggested implications for education planning based on all IEA studies conducted up to 1994.

Data collection

Some practical and theoretical considerations

The question of what kind of data should be collected in an international comparative assessment study cannot be answered unambiguously. The question is not a trivial one when one realizes that in most IEA studies more than twenty countries participate and more than forty countries participated in TIMSS. Many participants may differ in the functions or goals they want to realize through the study. Some may want to emphasize description of a small number of indicators, while others strive for a large number of variables to analyse their country's data properly. In addition, according to its mission, IEA wants to create opportunities to conduct cross-national analysis in order to enhance the understanding of the

functioning of education systems at all levels. There is always the dilemma between desirability and feasibility: researchers may desire to collect as much data as possible to be able to do in-depth secondary analyses of a rich database, while the usually restricted possibilities to collect data in schools as well as limited budgets put severe limitations on the size of the data collections. Therefore, in this type of study compromises have to be found among the interests of all participating countries. IEA is therefore striving for a design and for instruments that are as 'equally unfair' as possible to all participating countries.

In addition, for an effective and efficient study, a well-thought-out conceptual framework is necessary for addressing the issues to be included in a study. Almost all of the functions mentioned above need to measure education achievement and other outcomes of education on three levels of the education system.

| <i>Assessment of:</i> | <i>System level:</i> |
|--|----------------------|
| what students learn | micro |
| what and how schools and teachers teach | meso |
| what the community values (what students should learn) | macro |

IEA studies typically address all three levels by distinguishing three aspects of the curriculum:

- intended curriculum – what should be taught and learned, which is usually measured by analysing documents such as official syllabuses, course outlines and text books;
- implemented curriculum – what actually is being taught or taking place in schools and classroom (content, time allocations, instruction strategies etc.), which is usually measured through questionnaires (or observations); and
- attained curriculum – what students attain or learn in terms of cognitive skills, attitudes etc., which is usually measured

through tests. In the conceptual model for TIMSS, for example, the variables influencing education are seen as 'situated in a series of embedded contexts starting from the most global and moving to the most personal one', as is illustrated in Figure 2.

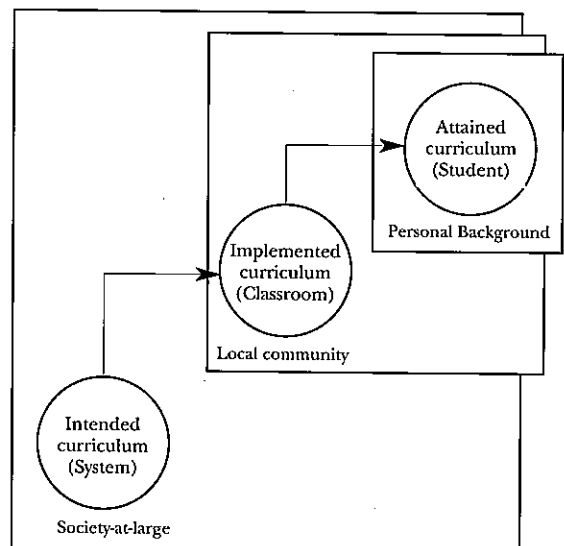
For more information about the conceptual approach of IEA, see Robitaille and Garden (1992) and Plomp (1992).

In a typical IEA study, many activities have to be completed to collect and provide data and indicators of good quality, such as curriculum analysis; instrument development (including pilot testing, translation etc.); sampling; production of instruments; data collection, cleaning and file building; quality control in participating countries of each component; data analysis; and report writing.

What data to collect: some examples

Within the practical and theoretical considerations previously discussed, the ques-

Figure 2: The conceptual framework for TIMSS



Source: D. F. Robitaille (ed.), *Curriculum Framework for Mathematics and Science*, pp. 26–7, Vancouver, Pacific Educational Press, 1993. (TIMSS Monograph 1.)

tions of what data should be collected in national and international assessment studies can be answered in various ways. Again, the answers depend on the functions as well as the research questions that the study is going to address. In addition, participating countries may want to use an international comparative study to find answers to some national questions as well. Therefore the 'what data' question has to be answered for each study separately. Here we will present some examples typical for IEA studies.

Data from what target populations?

The choice of the target population(s) is clearly a reflection of the (policy or research) questions in which one is interested. For example, in its cycle of achievement data collections, the OECD will collect data from 15/16-year-olds, to be able to provide policy-makers with a baseline profile of the achievement of students at (or close to) the end of compulsory schooling. On the other hand, in TIMSS, data have been collected for Grades 3 and 4 (population 1), Grades 7 and 8 (population 2) and the final year of secondary school (population 3), which allows for several comparisons as well as measurement of growth between two adjacent grades. By including common items in the tests for both populations, growth in mathematics and science from Grade 4 (elementary school) to Grade 8 (junior secondary school) can also be measured. In TIMSS, comparisons between populations 2 and 3 can be made as well. Moreover, the IEA target populations allow for monitoring the quality of education during compulsory schooling.

Multiple assessment measures

In TIMSS, achievement data have been collected in two ways. The achievement tests taken by all students in the study consisted

of open-ended questions and multiple-choice questions. Also, a sub-sample of the students in populations 1 and 2 took a series of performance assessment tasks in mathematics and science. The performance assessment, which was the same for both populations, was administered in a 'circus' format in which a student completed three to five tasks. The results are reported in Harmon et al. (1997). Table 3 presents some of the results of the performance assessment in combination with achievement results taken from Figure 1 and Table 1 for the countries that participated in the Grade-8 study in both the achievement testing and the performance assessment.

Table 3 illustrates the 'mirror' function of this descriptive data, which may lead to important questions for policy-makers and education practitioners in many countries. Thus, a number of countries have similar scores for all assessment measures; for example, Singapore is consistently above the international average and Spain, Portugal and Colombia are consistently below the international average. Interesting questions can be raised in, for example, the Netherlands and the Czech Republic. Both countries score well above the international average in mathematics and science achievement tests, but were only close to the international average in the performance tasks. If one values the capability of pupils to solve problems and performance tasks, then the satisfaction in these two countries about their high scores on the achievement tests should not overshadow the concerns they may have with their average results on the performance tasks. Some countries have one result that deviates from a pattern. For example, Switzerland is doing very well on the performance tasks and mathematics achievement, but average on science achievement. These examples illustrate that analysing the descriptive results on multiple assessment measures allows countries to raise questions that may lead to further, in-depth analyses and/or to discussions about the emphasis and focus in the curriculum.

Table 3: TIMSS Grade 8: Achievement and performance scores for mathematics and science

| <i>Mathematics</i> | | | | <i>Science</i> | | | |
|---|-----|--------------------------------------|----|---|-----|--------------------------------------|----|
| <i>Achievement test (scale pts)</i> | | <i>Performance tasks (av. %)</i> | | <i>Achievement test (scale pts)</i> | | <i>Performance tasks (av. %)</i> | |
| Singapore | 643 | Singapore | 70 | Singapore | 607 | Singapore | 72 |
| Czech Republic | 564 | Switzerland | 66 | Czech Republic | 574 | England | 71 |
| Switzerland | 545 | Australia | 66 | Netherlands | 560 | Switzerland | 65 |
| Netherlands | 541 | Romania | 66 | Slovenia | 560 | Scotland | 64 |
| Slovenia | 541 | Sweden | 65 | England | 552 | Sweden | 63 |
| Australia | 530 | Norway | 65 | Austria | 545 | Australia | 63 |
| Canada | 527 | England | 64 | Sweden | 535 | Czech Republic | 60 |
| Sweden | 519 | Slovenia | 64 | United States | 534 | Canada | 59 |
| New Zealand | 508 | Czech Rep. | 62 | Canada | 531 | Norway | 58 |
| England | 506 | Canada | 62 | Norway | 527 | New Zealand | 58 |
| Norway | 503 | New Zealand | 62 | New Zealand | 525 | Netherlands | 58 |
| United States | 502 | Netherlands | 62 | Switzerland | 522 | Slovenia | 58 |
| Scotland | 498 | Scotland | 61 | Scotland | 517 | Romania | 57 |
| Spain | 487 | Iran | 54 | Spain | 517 | United States | 55 |
| Romania | 482 | United States | 54 | Romania | 486 | Spain | 56 |
| Cyprus | 474 | Spain | 52 | Portugal | 480 | Iran, Islam. Rep. | 50 |
| Portugal | 454 | Portugal | 48 | Iran, Islam. Rep.) | 470 | Cyprus | 49 |
| Iran, Islam. Rep. | 428 | Cyprus | 44 | Cyprus | 463 | Portugal | 47 |
| Colombia | 385 | Colombia | 37 | Colombia | 411 | Colombia | 42 |
| Intl. Average | 513 | Intl. Average | 59 | Intl. Average | 516 | Intl. Average | 58 |

Sources: adapted from A. E. Beaton et al., *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS)*, Tables 1.1 and 1.2, Chestnut Hill, Mass., Center for the Study of Testing, Evaluation and Educational Policy, Boston College, 1996; adapted from A. E. Beaton et al., *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS)*, Figure 1.1, Chestnut Hill, Mass., Center for the Study of Testing, Evaluation and Educational Policy, Boston College, 1996; M. Harmon et al., *Performance Assessment: IEA's Third International Mathematics and Science Study*, Chestnut Hill, Mass., Center for the Study of Testing, Evaluation and Educational Policy, Boston College, 1997.

Background data

Background data are always collected in IEA studies (see Figure 2). Such data allow us to address research questions as to what factors contribute to good education. Another reason to collect such data is that they allow countries to search for determinants of national results in an international context.

In the IEA Reading Literacy study, Postlethwaite and Ross (1994) concluded that a large number of background variables influenced reading achievement. They were divided into several categories, namely indicators of student activities at home, school context, school characteris-

tics, school resources, school initiatives, school management and development, teacher characteristics, classroom conditions, teacher activities and teaching methods.

Postlethwaite and Ross (1994) analysed these indicators cross-nationally in the light of the question of what makes a school effective in reading. They found that in order to increase student reading performance, voluntary out-of-school reading should be fostered, particularly during the primary school years; schools should have classroom and/or school libraries; and teachers should emphasize reading for comprehension.

In general, the accumulated experiences gained in IEA studies in combination with the questions to be addressed in a study determine, to a large extent, what background data should be collected from schools, teachers and pupils.

Need for national assessment

International comparative studies can be utilized by a country to study its own education practice in an international comparative context. In the case of Switzerland, Moser (1997) analysed the mathematics data in TIMSS to determine the extent to which instructional practices (child-oriented versus subject-oriented instruction) and instructional variables (autonomy of students in child-oriented classes versus on-task behavior in subject-oriented classes) influenced learning outcomes. He looked not only at mathematics achievement but also at internal activity, self-activity and interest in mathematics. He concluded that instructional practices and instructional variables do not have a significant effect on mathematics achievement, but do have an effect on many other learning outcomes. In light of the much better results in Japan (a country with a high emphasis on subject-matter instructional practices and on on-task behaviour), he concludes that instructional practices in Switzerland can improve on these aspects.

Another example of a national analysis from Switzerland is related to our earlier conclusion that in TIMSS Switzerland is doing quite well on the performance tasks and on mathematics achievement, but average on science achievement. Ramseier (1997) analysed possible causes and concluded that this can be explained by a discrepancy between the Swiss science curriculum (teaching priorities) and the science test of the international study.

Most international comparative studies allow for a limited number of national questions ('national option'). The

example of Switzerland illustrates how important it is that countries participating in an international comparative study think beforehand about the national (policy and/or research) questions they want to address through such a study, and what typical characteristics of the national system need to be included in the background questionnaires to allow for relevant national analyses.

Concluding remarks

In the light of the above discussion and reflection on the significance of international comparative studies like those of IEA for evaluating and monitoring the quality of education, I offer the following general observations.

First, the relevance of participating in international comparative studies increases for a country if important reference countries participate as well. For that reason, a study like TIMSS has great relevance for the European Union, Northern America and a number of Asian countries. But for many countries, important reference countries are not restricted to the geographical region. Therefore, participation of Brazil and Chile in the TIMSS-repeat study can be of great importance for these countries, although they will, most probably, be the only participants from the Latin American region.

Second, IEA types of studies are logistically and methodologically complex studies. An important feature of IEA studies is the training of National Research Coordinators (NRCs). This is an essential component of the studies, as many NRCs may not be familiar with the methodology, and especially the specifics, of international comparative studies. A benefit of participating in such studies is the development of a network of researchers and specialists (in, for example, sampling, psychometrics, test development, data analysis etc.) which can be tapped when countries develop their own evaluation and national assessment studies.

Third, an important aspect, often overlooked, is the possibility of linking national assessments to international assessments. Proper linking of the two will not only increase the benefits a country can get from investments in assessment studies, but also be cost-efficient. Another cost aspect is related to the question of what data should be collected. As we illustrated in the examples given, policy and research questions should determine primarily what data should be collected. On the other hand, when cost factors have too much influence on what data will, or will not, be collected, one runs the risk of limiting the usability of the data collected. If IEA had collected only achievement data (which indeed allow for interesting indicators like those in Table 1 and Figure 1) in TIMSS, but had collected no data about schools, teachers and students, a country like Switzerland would never have been able to conduct national analyses in an international context and would have missed a unique opportunity to address some important national questions. It is often only a small increase in cost that makes the difference between collecting just achievement data or getting a rich dataset that allows for in-depth analyses of important issues.

References


- BEATON, A. E.; MULLIS, I. V. S.; MARTIN, M. O.; GONZALEZ E. J.; KELLY, D. J.; SMITH, T. A. 1996. *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, Mass., Centre for the Study of Testing, Evaluation and Educational Policy, Boston College.
- BEATON, A. E.; MARTIN, M. O.; MULLIS, I. V. S.; GONZALEZ E. J.; SMITH, T. A.; KELLY, D. L. 1996. *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, Mass., Centre for the Study of Testing, Evaluation and Educational Policy, Boston College.
- FOSHAY, A. W. 1962. *Educational Achievements of 13-year-olds in Twelve Countries*. Hamburg, UNESCO Institute of Education.
- HARMON, M.; SMITH, T. A.; MARTIN, M. O.; KELLY, D. L.; BEATON, A. E.; MULLIS, I. V. S.; GONZALEZ E. J.; ORPWOOD, G. 1997. *Performance Assessment IEA's Third International Mathematics and Science Study*. Chestnut Hill, Mass., Centre for the Study of Testing, Evaluation and Educational Policy, Boston College.
- HUSÉN, T.; POSTLETHWAITE, T. N. 1996. A Brief History of the International Association for the Evaluation of Educational Achievement (IEA). *Assessment in Education*, Vol. 3 (2), pp. 129–41.
- KEEVES, J. 1996. *The World of School Learning: Selected Findings from 35 Years of IEA Research*. Amsterdam, IEA.
- KELLAGHAN, T. 1996. IEA Studies and Educational Policy. *Assessment in Education*, Vol. 3, No. 2, pp. 143–60.
- MARTIN, M. O.; MULLIS, I. V. S.; BEATON, A. E.; GONZALEZ, E. J.; SMITH, T. A.; KELLY, D. L. 1997. *Science Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, Mass., Centre for the Study of Testing, Evaluation and Educational Policy, Boston College.
- MCKNIGHT, C. C.; CROSSWHITE, F.J.; DOSSEY, J.A.; KIFER, E.; SWAFFORD, J.O.; TRAVERS, K.J.; COONEY, T.J. 1989. *The Underachieving Curriculum: Assessing U.S. School Mathematics from an International Perspective*. Champaign, Ill., Stipes Publishing Company.
- MOSER, U. P. 1997. *Swiss Analysis of the TIMSS Data*. (Paper presented at the annual conference of the American Educational Research Association, Chicago, Ill.)
- MULLIS, I. V. S.; MARTIN, M. O.; BEATON, A. E.; GONZALEZ, E. J.; KELLY, D. J.; SMITH, T. A. 1997. *Mathematics Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, Mass., Centre for the Study of Testing, Evaluation and Educational Policy, Boston College.
- MULLIS, I. V. S.; MARTIN, M. O.; BEATON, A.

- E.; GONZALEZ, E. J.; KELLY, D. J.; SMITH, T. A. 1998. *Mathematics and Science Achievement in the Final Year of Secondary School: IEA's Third International Mathematics and Science Study*, Chestnut Hill, Mass., Boston College.
- OECD. 1997. *Education at a Glance: OECD Indicators*. (5th ed.). Paris, OECD/CERI.
- PLOMP, T. 1992. Conceptualizing a Comparative Educational Research Framework. *Prospects*, Vol. XXII, No. 3, pp. 278-88.
- . 1997. International Educational Research: The Case of IEA. In: S. Hegarty (ed.), *The Role of Research in Mature Education Systems*. Slough (England), National Foundation for Educational Research, pp. 184-95.
- POSTLETHWAITE, T. N.; ROSS, K. 1994. *Effective Schools in Reading: Implications for Educational Planners*. Amsterdam, IEA.
- RAMSEIER, E. 1997. *Task Characteristics and Task Difficulty: Analysis of Typical Features in the Swiss Performance in TIMSS*. (Paper presented at the European Conference on Educational Research, Frankfurt.)
- ROBITAILLE, D. F. (ed.). 1993. *Curriculum Frameworks for Mathematics and Science*. Vancouver, Pacific Educational Press. (TIMSS Monograph 1.)
- ROBITAILLE, D. F.; GARDEN, R. (eds.). 1989. *The IEA Study of Mathematics II: Contexts and Outcomes of School Mathematics*. Oxford, Pergamon Press.



International assessments: the United States TIMSS experience

Albert E. Beaton



Introduction

TIMSS is the largest international assessment ever done. It was done under the auspices of IEA and was directed from its International Study Center at Boston College. The assessment required the co-operation and co-ordination of education researchers, mathematics and science specialists, policy-makers, psychometricians and many others from the participating countries around the world.

Not surprisingly, the main findings of TIMSS are that there are many large differences in the way schooling is organized and in the way that students perform in different participating countries. Its reports have generated discussions in many parliaments and education ministries, as well as in the popular media. In the United States, some results have been reported by the President and also mentioned in his annual State of the Union address. The Secretary of Education and the Director of the National Science Foundation have cited TIMSS results. The mass media have disseminated the results widely. So far, the TIMSS discussions have focused mainly on education output and detailed research into the correlates of this output is yet to be done.

Before proceeding, it is worth noting the enormous size of TIMSS. It assessed five grades (Grades 3, 4, 7, 8 and 12 in the United States) in two subject areas (mathematics and science). Over forty countries and half a million students participated. The TIMSS tests included short or extended written responses to questions as well as the selection of multiple-choice options; some students were asked to design and perform hands-on experiments using laboratory equipment. Questionnaires were administered to students, teachers and principals. Data on national curricula, textbooks, organization and control were also collected. The amount of information collected is enormous.

The assessment is now done; the data scored, cleaned and analysed; and TIMSS has already produced a series of reports that have been widely cited in the United States and around the world. The reports cover mathematics and science achievement at the primary school level (Mullis et al., 1997; Martin et al., 1997), the middle school level (Beaton, 1996; Beaton et al., 1996), and at the end of secondary school (Mullis et al., 1998). The results of the performance assessments have been published (Harmon et al., 1997). These reports contain a wealth of information

about student performance, student backgrounds and attitudes, their teachers and their schools. TIMSS has also published qualitative information about the participating school systems (Robitaille, 1966). Two volumes have been published on curricula intentions in mathematics and science (Schmidt et al., 1997). Articles for journals have been published and are in preparation.

It is important that the results of such a study be credible and TIMSS used the most advanced assessment technology, with great attention to quality. The sampling, test development, administration, scoring, database construction, analysis and reporting were all carefully devised and controlled, with quality checks at every step of the way. The technology has already been documented in a report on quality assurance procedures (Martin and Mullis, 1996) and in two technical reports (Martin and Kelly, 1996; 1997), and more documentation will be forthcoming in a third technical report, which is now in preparation. The TIMSS reports, data, and other information are available on the World Wide Web at <http://www.csteep.bc.edu/timss>.

This paper discusses some of the results for the United States, particularly for TIMSS population 3, which covers students at the end of secondary schooling, Grade 12 in the United States. Participation at this level was not compulsory, with the result that twenty-four countries participated. There were three different tests at this level: a general mathematics and science literacy test for all students; an advanced mathematics test for students taking advanced mathematics; and a physics test for those taking that subject. The United States participated in all three parts, although not all other countries did.

United States results

The results for United States secondary school seniors were very disappointing (Table 1). The mathematics and science lit-

eracy test was designed to see whether students had the ability to solve the sort of problems that adults encounter at work and in everyday life. The questions were not tied to a secondary school curriculum, and the students would have met the necessary mathematical and scientific knowledge and processes earlier in their education, perhaps several times. The students finished significantly below the international average in this test. This result is especially disconcerting because United States students were above average in the primary-school assessment and at around the average at the middle-school level.

The results for those students taking advanced mathematics courses were even worse. The results are shown in Table 2. The United States again scored below the international average, coming out second last, outperforming only Austria. The mathematics test was very difficult and contained a number of calculus questions. Only a few United States advanced mathematics students had taken calculus, but this was also true in some other countries. Curiously, the United States did comparatively less well in geometry than in calculus.

The results for United States students in physics were no better; they were dead last among the sixteen countries that participated in the physics test (Table 3).

The results were a surprise to me; I expected the United States to do better. I believed that it would be somewhere in the middle, but not close to being the best in the world in the year 2000. My first instinct was to check the results in case there was some data-processing blunder, and so each calculation was checked and checked again. A review of all procedures convinced me of the accuracy of the results.

I knew that these results would be bad news in the United States and I knew what often happens to messengers who bring bad news. I knew that TIMSS would be subjected to extraordinary scrutiny, looking for a way to discredit the results. I welcome this attention, since we reviewed past criticisms of international studies and did

Table 3: Distribution of physics achievement for students in their final year of secondary school

| Country | Mean achievement |
|---|------------------|
| <i>Significantly higher than international average</i> | |
| Norway | 581 |
| Sweden | 573 |
| Russian Federation | 545 |
| Denmark | 534 |
| <i>Not significantly different from international average</i> | |
| Slovenia | 523 |
| Germany | 522 |
| Australia | 518 |
| Cyprus | 494 |
| Latvia* | 488 |
| Greece | 486 |
| <i>Significantly lower than international average</i> | |
| Switzerland | 488 |
| Canada | 485 |
| France | 466 |
| Czech Republic | 451 |
| Austria | 435 |
| United States | 423 |
| International average | 501 |

* Latvian-speaking schools.

Source: adapted from I. V. S. Mullis et al., *Mathematics and Science Achievement in the Final Year of Secondary School: IEA's Third International Mathematics and Science Study (TIMSS)*, Table 8.1. Chestnut Hill, Mass., Center for the Study of Testing, Evaluation, and Educational Policy, Boston College, 1998.

in the primary and middle school years and yet had the largest class size by far, averaging over forty students per classroom at Grade 8. The use of calculators did not correlate well with student performance, nor did the length of the school year or the amount of instruction time or the amount of homework. However, some TIMSS videotapes showed differences in the way instruction was delivered. Starting school at a younger age is also problematic because Scandinavian students start school a year later and yet catch up by the time they graduate from secondary school. I guess that if there were a simple answer we would know it by now.

Is TIMSS unfairly comparing its general students in the United States against

Table 4: TIMSS coverage index (TCI)

| Country | % |
|--------------------|------|
| Slovenia | 87.8 |
| Norway | 84.0 |
| France | 83.9 |
| Switzerland | 81.9 |
| Netherlands | 78.0 |
| Czech Republic | 77.6 |
| Austria | 75.9 |
| Germany | 75.3 |
| Sweden | 70.6 |
| New Zealand | 70.5 |
| Canada | 70.3 |
| Australia | 68.1 |
| Hungary | 65.3 |
| United States | 63.1 |
| Denmark | 57.7 |
| Iceland | 54.5 |
| Italy | 51.5 |
| South Africa | 48.9 |
| Russian Federation | 48.1 |
| Cyprus | 47.9 |
| Lithuania | 42.5 |

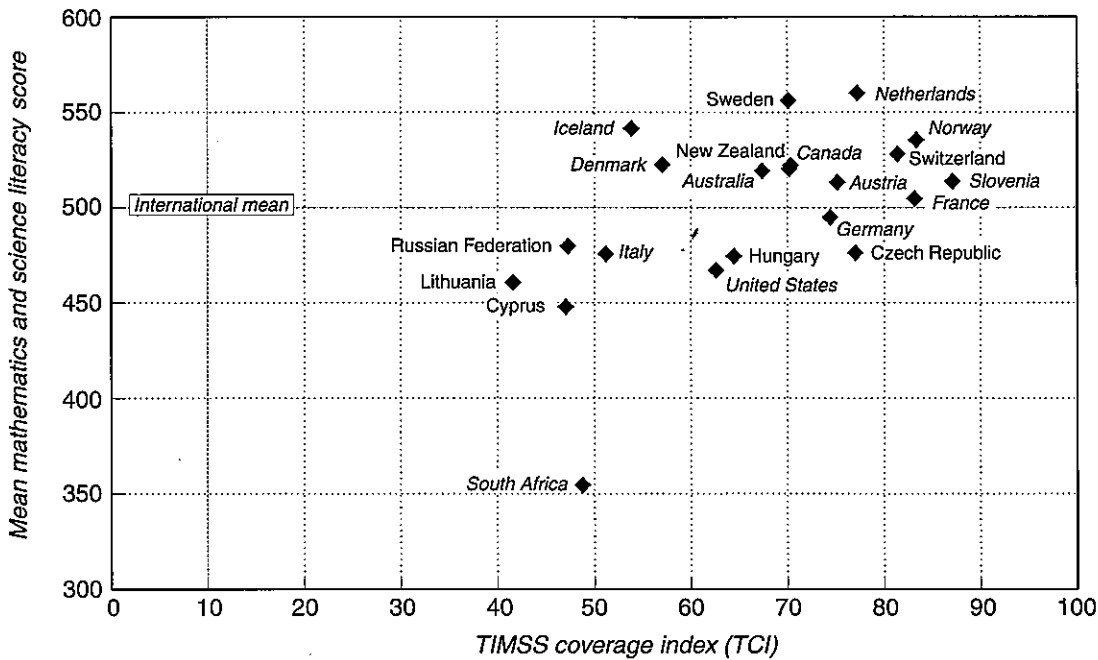
Source: adapted from I. V. S. Mullis et al., *Mathematics and Science Achievement in the Final Year of Secondary School: IEA's Third International Mathematics and Science Study (TIMSS)*, Table 2. Chestnut Hill, Mass., Center for the Study of Testing, Evaluation, and Educational Policy, Boston College, 1998.

Comments on the TIMSS results

The TIMSS results make it clear that there are no easy answers to the question of why some countries perform better than others, and there are no easy fixes. Naturally, we looked first at the manipulable policy variables and they did not show consistent results. For example, many believe that reducing class size will improve performance, but Korean students did very well

Albert E. Beaton

Figure 1: Mean mathematics and science literacy achievement by TIMSS coverage index for students in their final year of secondary school



Countries shown in italics did not satisfy one or more guidelines for sample participation rates or student sampling.

Source: adapted from I. V. S. Mullis et al., *Mathematics and Science Achievement in the Final Year of Secondary School: IEA's Third International Mathematics and Science Study (TIMSS)*, Figure 1.2. Chestnut Hill, Mass., Center for the Study of Testing, Evaluation, and Educational Policy, Boston College, 1998.

the elite students of other countries? It is a common error to believe that the United States educates all of its students but that many other countries educate only their elite or best students. If this were so, it would be problematic to compare ordinary students in the United States with the best students elsewhere. However, it is not so.

This problem was real to some degree in the 1960s and before, but the European education system has changed and many countries have caught up with the United States. To investigate the problem, TIMSS created a Test Coverage Index (TCI), which is the ratio of the population size, as estimated from the TIMSS sample, to the number of students at the appropriate age levels, as taken from official documents such as a census. These statistics were carefully scrutinized to assure that the TIMSS samples were representative of the national populations. The numerator is the estimated number of students and the

denominator is number of students who are in school, plus those students excluded from the sample because of various conditions (IEP and LEP students in the United States) and students who have dropped out of school.

Table 4 shows the TCI index for the twenty-one countries that participated in the mathematics and science literacy tests. Slovenia's sample covers the largest proportion of students (88 per cent), and samples from Norway, France and Switzerland also cover more than 80 per cent of their age cohorts. The United States sample covered only 63 per cent of the age cohort, less than most other countries. The samples of four countries covered less than 50 per cent of the youth group. From these data, we conclude that the poor standing of the United States is not due to its comparison to elite student groups; if anything, a larger percentage of students in many other countries is attending school.

Table 5: Average age of students in their final year of secondary school

| Country | Years |
|--------------------|-------|
| Iceland | 21.2 |
| South Africa | 20.1 |
| Switzerland | 19.8 |
| Norway | 19.5 |
| Germany | 19.5 |
| Austria | 19.1 |
| Denmark | 19.1 |
| Sweden | 18.9 |
| France | 18.8 |
| Slovenia | 18.8 |
| Italy | 18.7 |
| Canada | 18.6 |
| Netherlands | 18.5 |
| United States | 18.1 |
| Lithuania | 18.1 |
| Czech Republic | 17.8 |
| Australia | 17.7 |
| Cyprus | 17.7 |
| New Zealand | 17.6 |
| Hungary | 17.5 |
| Russian Federation | 16.9 |

Source: adapted from I. V. S. Mullis et al., *Mathematics and Science Achievement in the Final Year of Secondary School: IEA's Third International Mathematics and Science Study (TIMSS)*, Table 1.1. Chestnut Hills, Mass., Center for the Study of Testing, Evaluation, and Educational Policy, Boston College, 1998.

an inclusive school system with high performing students.

Age of students

TIMSS found substantial variation in the age at which students in different countries completed secondary school, as seen in Table 5. The average age of students varied from 21.2 years (Iceland) to 16.9 years (the Russian Federation). The average of United States students was 18.1 years. These differences can be explained in part by school policies. As mentioned above, Scandinavian students begin schooling a year later than United States students and so are a year older when they finish their twelve grades. Iceland not only starts a year later but also has a fourteen-year education system. The Russian Federation has an eleven-year system, so its students tend to be a year younger.

And yet, age does not seem to be a clear determinant of student proficiency. Although Iceland did quite well on the test, the South African students were the second oldest and had the lowest performance. The Australian students averaged somewhat younger (17.7 years) than the United States students but significantly outperformed them. Based on these data, changing the ages at which students attend school does not seem likely to affect performance very much.

Course-taking

The TIMSS data suggest that course-taking practices are important, but again the data do not point to simple causal explanations. Some 34 per cent of United States secondary school seniors do not take mathematics courses; however, even smaller percentages of students are enrolled in mathematics courses in Canada, Iceland, the Netherlands and Switzerland, and yet the students in these countries outperformed the United States students. Some 53 per cent of United States students are not enrolled in science courses, but many

It is interesting to note that the selectivity of the school systems in TIMSS is not closely related to the performance of its students. Figure 1 is a scattergram showing the relationship between TCI and average student performance. If selectivity were an important factor, we would expect selective countries to be in the upper left-hand corner – those with low TCIs would have high average achievement – and expect unselective countries to be in the lower right-hand corner – those with high TCIs would have low performance. But this is not the case; in fact there is a slight tendency for the countries with high TCIs to have high performing students as well. The countries with low TCIs tended to do very poorly on the tests. It is, therefore, possible to have

Albert E. Beaton

Table 6: Gender differences in physics achievement for students having taken physics in their final year of secondary school

| Country | Males | | Females | | Gender difference | PTCI ¹ (%) |
|-----------------------|------------------------|------------------|------------------------|------------------|-------------------|-----------------------|
| | Percentage of students | Mean achievement | Percentage of students | Mean achievement | | |
| France | 61 (2.0) | 478 (4.2) | 39 (2.0) | 450 (5.6) | 28 (7.0) | 20 |
| Cyprus | 63 (2.5) | 509 (8.9) | 37 (2.5) | 470 (7.1) | 40 (11.4) | 9 |
| Latvia ² | 51 (3.7) | 509 (19.0) | 49 (3.7) | 467 (22.6) | 42 (29.5) | 3 |
| Canada | 57 (3.2) | 506 (6.0) | 43 (3.2) | 459 (6.3) | 47 (8.7) | 14 |
| Norway | 74 (1.8) | 594 (6.3) | 26 (1.8) | 544 (9.3) | 51 (11.2) | 8 |
| Sweden | 67 (3.4) | 589 (5.1) | 33 (3.4) | 540 (5.3) | 49 (7.4) | 16 |
| Russian Fed. | 54 (2.0) | 575 (9.9) | 46 (2.0) | 509 (15.3) | 66 (18.2) | 2 |
| Czech Rep. | 38 (2.4) | 503 (8.8) | 62 (2.4) | 419 (3.9) | 83 (9.7) | 11 |
| Switzerland | 51 (1.8) | 529 (5.2) | 49 (1.8) | 446 (3.6) | 83 (6.3) | 14 |
| Greece | 68 (2.1) | 495 (6.1) | 32 (2.1) | 468 (8.1) | 28 (10.1) | 10 |
| Germany | 69 (3.0) | 542 (14.3) | 31 (3.0) | 479 (9.1) | 64 (17.0) | 8 |
| Australia | 66 (3.8) | 532 (6.7) | 34 (3.8) | 490 (8.4) | 42 (10.8) | 13 |
| Austria | 38 (3.5) | 479 (8.1) | 62 (3.5) | 408 (7.4) | 71 (11.0) | 33 |
| United States | 52 (2.4) | 439 (4.3) | 48 (2.4) | 405 (3.1) | 33 (5.3) | 14 |
| Denmark | 80 (2.3) | 542 (5.2) | 20 (2.3) | 500 (8.1) | 42 (9.6) | 3 |
| Slovenia | 72 (3.7) | 546 (16.3) | 28 (3.7) | 455 (18.7) | 91 (24.8) | 39 |
| International average | | 523 | | 469 | 54 | |

1. The Physics TIMSS Coverage Index (PTCI) is an estimate of the percentage of the school-leaving age cohort covered by the TIMSS final-year physics student sample.

2. Latvian-speaking schools.

() Standard errors appear in parentheses. Because results are rounded to the nearest whole number, some differences may appear inconsistent.

Source: adapted from I. V. S. Mullis et al., *Mathematics and Science Achievement in the Final Year of Secondary School: IEA's Third International Mathematics and Science Study (TIMSS)*, Table 8.4. Chestnut Hills, Mass., Center for the Study of Testing, Evaluation, and Educational Policy, Boston College, 1998.

other countries have similar enrolment numbers. Only 14 per cent of United States students take advanced mathematics or physics. The curricula of different countries vary, however, and it is worth noting that the high-scoring Norwegian physics students study the subject for three years, compared to students in the United States and many other countries, who only study physics for one year.

Outside activities

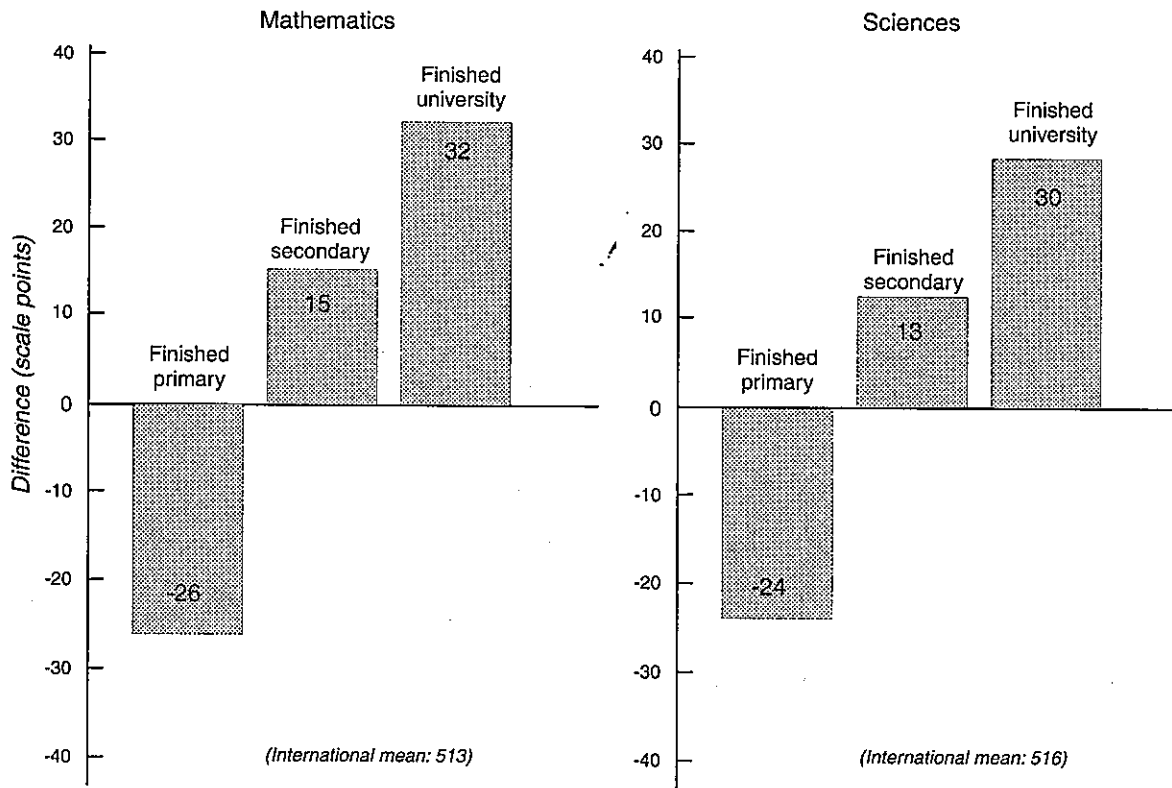
In many ways, students around the world are similar; they all watch television, chat with friends, and behave generally like the

teenagers in other countries. One big difference is that United States students tend to have jobs outside of school – in fact, over half those who responded indicated that they worked several hours a day. No other group of students approaches this level of work. This finding is worth more study.

Gender differences

At this education level and in all countries, males outperform females in mathematics and science literacy, advanced mathematics, and physics, although the gender gap is occasionally not statistically significant. United States males significantly outper-

Figure 2: Parents' education: achievement differences from international mean



formed United States females in all tests at this level. These boys also tended to like mathematics and science better than their female peers did. The gender differences in physics are shown in Table 6. What is interesting to note is that the percentage of United States physics students who are male (52 per cent) is very close to that of females (48 per cent), whereas in high-scoring Norway the percentage of males is 74 per cent and that of females is 26 per cent. I did a simple calculation to standardize the averages in order to remove the effect of differential proportions of males and females and found that such an adjustment would not substantially affect the results in the table. The gender differences are more pronounced at the secondary level than at the primary- and middle- school levels.

Home environment

One factor is consistent over all TIMSS tests at all grade levels. In each participating

country, the children of university-educated parents on the average outperform the children of parents with secondary school education, and children of parents whose education ended with secondary school in turn outperform the children whose parents did not complete secondary school. It is also true that students who have educationally rich homes with many books, calculators and study space do better in school than those who do not. The average advantage over all countries for the eighth grade is shown in Figure 2. How parents affect the learning of their children and the teaching in their schools needs considerable further study.

Curriculum

Bill Schmidt of Michigan State University has been investigating the curricula of various countries. He concludes that the United States curriculum – if there is one – is a 'mile wide and an inch deep'. He notes

that United States textbooks are very large and cover many topics, many more than are covered by the textbooks of other countries. Also, in the United States curriculum, many topics are covered again and again, instead of being covered once. He concludes that a more focused curriculum would be helpful.

General comments

From TIMSS, it is clear that the students in the United States secondary schools are not performing as well in mathematics and science as secondary-school leavers in most other industrial countries. The TIMSS data suggest no simple reasons for this phenomenon and no simple solution is likely. It is clear that the United States is currently doing very well economically, even with this deficit. It seems to me that we have to think about the consequences of this deficit and how much we want to remove it.

From the discussions that have followed the TIMSS reports, we seem to want to take the deficit very seriously. In the past, two Presidents have taken the position that the United States should be number one in mathematics and science by the year 2000. Although this intention will not be attained, it is still our intention. There are currently many education reform efforts and most of them seem likely to be able to make a small dent in the performance deficit. But none seems likely to result in fully closing this gap. I think that we have to stand back and ask ourselves why.

I think that the United States school system is quite responsive to the social demands that are placed on it. For example, most of the public would like schools to demand higher standards of achievement and TIMSS has shown that the high standards proposed by the National Assessment Governing Board are attained by many students in other countries. However, when such standards result in many student failures, there is a public outcry against the standards and the consequences

for the students who fail. The school has to move carefully to do the best it can with such competing demands.

The schools have been concerned about student self-concepts and self-esteem, and apparently have succeeded. United States students think that they are doing well in mathematics and science, although they are not. The schools have been teaching the students about various cultures and their similarities and differences, but not about calculus and physics. Most United States students have devoted a large amount of time to sex education, but sex education was not covered in the TIMSS tests. We have placed many demands on the school system and it seems to me that it has responded reasonably well to those demands.

I think that we now need to reconsider what we want the schools to do and to set our priorities. If we want to be number one in the world in mathematics and science, we will have to reorganize our schools and their staffing drastically and bear the expense. But do we really want that enough? If the cost is too great – socially or economically – we will have to decide what we do want and set priorities. We need a reasonable consensus on what the school should produce and fund it accordingly.

References

- BEATON, A. E. 1996., *Mathematics in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, Mass., Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- BEATON, A. E.; MULLIS, I. V. S.; MARTIN, M. O.; GONZALEZ, E. J.; KELLY, D. L.; SMITH, T. A. 1996. *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, Mass., Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- MARTIN, M. O.; KELLY, D. L. 1996. *Third Inter-*

- national Mathematics and Science Study Technical Report*. Chestnut Hill, Mass., Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- MARTIN, M. O.; MULLIS, I. V. S. 1996. *Quality Assurance in Data Collection*. Chestnut Hill, Mass., Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- MARTIN, M. O.; MULLIS, I. V. S.; BEATON, A. E.; GONZALEZ, E. J.; KELLY, D. L.; SMITH, T. A. 1997. *Science Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, Mass., TIMSS International Study Center, Boston College.
- MULLIS, I. V. S.; MARTIN, M. O.; BEATON, A. E.; GONZALEZ, E. J.; KELLY, D. L.; SMITH, T. A. 1997. *Mathematics Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, Mass., TIMSS International Study Center, Boston College.
- 1998. *Mathematics and Science Achievement in the Final Year of Secondary School: IEA's Third International Mathematics and Science Study*. Chestnut Hill, Mass., Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
- ROBITAILLE, D. F. 1997. *National Contexts for Mathematics and Science Education: IAE's Third International Mathematics and Science Encyclopedia of the Educational Systems Participating in TIMSS*. Vancouver, B.C., Pacific Educational Press.
- SCHMIDT, W. H.; MCKNIGHT, C. C.; RAIZEN, S. A. 1997. *Splintered Vision: an Investigation of United States Science and Mathematics Education*. 2 volumes. Dordrecht, The Netherlands, Kluwer Academic Publishers.

Comments on 'International comparative assessments' and 'The TIMSS experience'

Jahja Umar

Introduction

I am pleased to have the opportunity to comment on these excellent papers. I will first address several issues considered by Plomp. Two purposes of international comparative achievement studies are stated as:

- to provide policy-makers and education practitioners with information about the quality of their education in relation to relevant reference groups; and
- to assist in understanding the reasons for observed differences between education systems.

Jahja Umar

Issues for policies

I agree that policy-makers can expect help in generating policy questions from an international comparative study, but I think such a study should also help them answer some policy questions, especially as regards policies that are common to some countries. In other words, the policy-makers might expect more on comparisons in relation to relevant common policies rather than to reference groups in a table of ranks. This would require different methods of data analysis and presentation of the results. It

Also, extending the comparisons to some external variables using methodology such as in Muthen (1986, 1988) might result in better control for variables such as OTL, age, class size, courses and so forth. If external variables include some important policy variables, more information can be provided to the policy-makers.

References

- MUTHEN, B. 1986. *Instructionally Sensitive Psychometrics: Applying Structural Models to Educational Achievement Data*. University of California at Los Angeles.
- . 1988. Some Uses of Structural Equation Modeling in Validity Studies: Extending IRT to External Variables. In: H. Wainer and H. I. Braun (eds.), *Test Validity*. Lawrence Erlbaum Associates: Hillsdale, N. J.

Comments on 'International comparative assessments' and 'The TIMSS experience'

Giray Berberoglu

Introduction

The paper presented by Plomp addresses two important purposes of international comparisons:

- to provide policy-makers and educators with information about the quality of education in relation to the reference groups; and
- to assist in understanding the reasons for observed differences among education systems.

Problems of comparison

In line with the first purpose, which is called 'mirror function', countries should choose a reference country. However, no country seems to be happy with the results. Which country is the model to imitate? The United States is the model for most developed and developing countries. However, the United

States results are far below the expected level. The standards set by the country may be more helpful in deciding the quality of their education system. This is a criterion-referenced decision rather than a norm-referenced one.

Interpreting differences

On the other hand, differences among countries may reflect curriculum differences rather than differences observed in student achievement. In this respect, tests used by the TIMSS study should emphasize higher order thinking skills rather than the subject-matter domain of the curricula. This would bring more fairness to the comparison process. Indicating whether students are able to interpret a given graph is more meaningful than indicating how much they know about specific subject matter. As I will explain later, curricula used by different countries may make the tests unequivalent

Giray Berberoglu

across countries and/or languages. Understanding the reasons for observed differences among education systems requires more elaborate statistical analyses within each country.

The paper presented by Beaton addresses possible reasons for low mean scores among American students.

As mentioned previously in the discussion, it is not possible to cover the subject matter of every curriculum used by different countries. Tests should sample cognitive processes rather than subject matter. It is also very important to consider other intervening variables besides the curricula, such as the age of the students, course-taking practices, outside activities and so on, in interpreting the results.

■ ■ ■ Analysis of results ■ ■ ■

More elaborate statistical analysis, such as multivariate analysis, may help us to better understand the TIMSS data. For example, there is a need for analysis that will control for the effects of some variables on the achievement scores. Combining questionnaire data with achievement scores will help us to understand the factors affecting student achievement. Caution is required. For example, in a question regarding reaction to science as a subject, 71 per cent of United States students responded that they 'liked it' or 'liked it a lot'. However, countries where students did better on the science test produced a lower percentage for that particular response: Hungary, 62 per cent; Japan and Germany, 59 per cent. These results prompted Myron Aitkin and Paul Black to say that in order to achieve higher scores on the tests, students in the United States will need to be taught in such a way that they will like the subject less than they do at present!

Another possibility for obtaining more information from TIMSS is to evaluate subcomponents of the tests rather than reporting a single test score. For instance, males outperform females in math, but it

would be interesting to see what happens if word problems, computation, and geometry items were evaluated separately across groups and countries.

■ ■ ■ Test translation problems ■ ■ ■

One technical problem that should be given special emphasis is test translation. Improper translations may make the test instruments easier or more difficult for students in some countries. Some item formats may be inappropriate for some language structures. For a fair and valid comparison among countries, special care must be given to the test translation process.

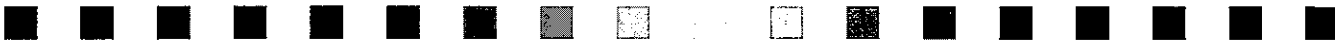
In a study carried out by Hambleton and myself of about twenty participating countries/territories surveyed regarding the test translation processes, several reported using no modification in translating the tests in line with the *TIMSS Survey Operations Manual*. Others reported some modifications. There were also reported problems related to the qualifications of translators. Almost all of the participants reported certain problems in terms of language equivalence and difficulties in finding out the equivalent terms and terminology in the translation.

Differences in the structure of the languages created some serious translation problems, especially in passive sentences in Republic of Korea, Spain and Sweden. Long sentence structures also created translation problems. For instance, France and Sweden preferred to use shorter sentences. It is interesting that France and Sweden had fewer problems in terms of reading ability than English-speaking participants because students' reading ability interfered less with the test scores for them. Even those that used the original English version (Australia, Ireland, New Zealand, Scotland and Singapore) reported some problems. They made some cultural adaptations, especially in the questionnaire. These modifications clearly violated the principles of translation proposed in the

TIMSS *Main Study Manuals*. Non-equivalent forms generated by the participating countries may be jeopardizing comparisons. More can and should be done to improve the translation process itself.


These are just a few of the issues we need to continue to discuss and research, as it is clear that TIMSS and other international comparative assessments are certain to become part of the international education landscape.





7. Overview and synthesis: the role of measurement and evaluation in education policy

Edmund W. Gordon



■ Introduction

According to Dr Messick, the key issues of concern in the discussion at the round-table were:

- low-stakes versus high-stakes assessment and the attendant issues of equity;
- current policy uses of student assessments; and
- large-scale assessment as policy research.

These issues were revisited through-out the discussion, but it was the issues of equity and fairness that dominated the entire discussion. This ubiquitous concern in the discussion for fairness and equity was anticipated by Dr Messick, who argued that from student assessments conducted to better inform classroom decisions about students; through large-scale assessments as the basis for public policy decision-making; to individual assessments for the purposes of admission, promotion, graduation and professional licensure, the importance of fairness and equity issues in education assessment can not be over emphasized. In his opening remarks he stated, 'They (equity and fairness issues) permeate all the topics we will discuss'.

Issue 1: Low-stakes versus high-stakes education assessment and the attendant issues of equity

Low-stakes education assessments provide information about student performance to educators and policy-makers, with no rewards or sanctions attached to the quality of the performance. High-stakes education assessments provide information about performance to which rewards and sanctions are attached. As those rewards and sanctions gain in consequential significance for students, for professional educators or for society, questions of fairness and equity become more salient. The confluence of issues concerning equity and assessment confront the field of education assessment with very difficult problems.

Dr Messick argued that 'it is difficult', if not impossible, for assessment to be fair to individuals in terms of equity; to groups in terms of parity or absence of adverse impact; to institutions in terms of efficiency; to societies in terms of costs and benefits; all at the same time. Each of these is a serious equity problem in assessment. If you have to do them all at once, how can balances and the trade-off's be treated? This is not necessarily an assessment problem. It

is a policy problem, but it is a policy problem that affects the way assessment will play itself out in the next several years.

The concern for equity is a public policy problem that also affects the context in which education assessment must operate. Those of us concerned with measurement find ourselves mired in the tensions surrounding issues reflected in differing conceptions of equality and equity, with respect to educational treatments, to educational outcomes and to the differential assessment of both. Traditional concepts of equality have emphasized sameness of input and outcome, while modern concepts have moved to a concern for equity as is reflected in appropriateness and sufficiency of opportunities to learn, in the quality of achievement and in the capacity of the assessment to demonstrate accurately what each test-taker knows and can do. In these concerns it is obvious that psychometric issues are being combined with policy issues. In the round-table discussion attention was given to what can be done about these equity issues. It has been suggested that adjustments could be made in test scores, that extra points could be given to people who have certain disadvantages or that group differences could be taken into account in the selection of test items. The problem with all of these adjustments is that they run the risk of biasing the measure of the construct and the nature of test-score differences. This could result in biasing the construct with respect to all other comparisons. However, consensus holds that the critical factor in measurement should be to maximize the validity of construct interpretation. When we bias the construct measures, we are corrupting the indicator, which threatens the scientific uses of the construct measure. If there are adjustments to be made, they should be made in terms of the actions that follow construct measurement as a matter of policy and not actions that modify the construct.

Stimulated by the paper on Equity in Education and Assessment by Caroline Gipps, the discussion of accommo-

dations and adaptations based on the social divisions (class, gender, race, etc.) to which learners and examinees may be assigned included a wide range of concerns. Among these were the following three.

Firstly, what are the political, pedagogical and psychometric aspects of the several categories by which we group learners? It was argued that the political aspects adhere to the status assigned to each group, which, in turn, influences access, expectation, opportunity and reward. The pedagogical aspects adhere to the differences in functional characteristics, which may be associated with group identity or membership, but may also be a matter of individual variation. The psychometric aspects are less clear, but may adhere to the affective and attributional factors that could come to be associated with perceived functions of testing and their differential impacts on members of specific groups, as in Steele's 'fear of stereotype confirmation'. Despite the ambiguities surrounding the relevance of social divisions into which persons may be classified, attention was called to the continuing dominance of these categories in the analysis and reporting of education assessment data. It was generally agreed that there may be more pedagogically relevant categories, but the weight of tradition and political sensitivity suggest that existing categories like class, gender, national origin and race will continue to be the dominant indicators of group membership for purposes of data analysis and reporting.

Secondly, at the core of the pedagogical problem, which is sometimes confused with the psychometric problem, is the persistent association between academic achievement and membership in one or more of the human social divisions. The separation of one's school achievement from one's social status is the challenge identified by James Coleman in his 1965 study *Equal Educational Opportunity*. In that work, Coleman argued that the educational challenge to the nation was to uncouple academic achievement from membership in

any specific social division. One's race or class can not reliably predict one's academic achievement. In the round-table discussion it was argued that when the quality of educational treatment accurately reflects the specific outcomes to be measured, this circumstance tends to reduce the issue of fairness in education assessment. When educational treatments are appropriate and sufficient, the problems of equity in education assessment are greatly reduced.

Finally, while it is not primarily a psychometric problem, the absence of appropriateness and sufficiency of educational treatments does present problems for educational measurement. If assessment procedures are to meet the criteria for equity, it is argued that opportunity to learn as well as opportunity to demonstrate what one has learned must be addressed as prior conditions of measurement. The discussion of specifications of criteria for judging opportunity to learn did not progress very far during the round table; however, implications for measurement of the extent to which students benefited from assumed opportunities received some attention in the papers by Beaton, Everson, Gipps and McGaw. It was the sense of the discussion that, given the current level of psychometric knowledge and technology, the opportunity to learn issue can only be addressed through causal inferences drawn from the interpretation of achievement test data. One possible exception to this conclusion is the conjoining of teaching, learning and assessment, as in the Advanced Placement and New Standards programmes in the United States. In both of these programmes instruction and assessment are closely linked to agreed-upon standards. The emerging standards movement in the United States may have the capacity to move the field of education assessment in that direction.

Issue 2: Current policy uses of student assessments

Issues concerning low-stakes and high-stakes assessment are not limited to implications

for equity and fairness. Questions related to the low-stakes/high-stakes consequences of using education assessment data are highlighted in the new emphasis on standards as a current policy issue in education, as well as in the uses of student education achievement data. Test performance and student achievement data are used for a wide variety of purposes, for example, to inform admissions, placement and instructional decisions; to provide information concerning the functioning of education systems; to certify specific levels of student achievement; and to hold students, educators and schools accountable for the quality of student achievement. Central to all of these purposes is agreement on the ends toward which students, educators and systems work. The concern for education standards is but a reflection of the centrality to the education enterprise of what it is that we want students to know, to be able to do and, as persons, to be.

As Everson discussed in his paper on Education Standards, the modern standards movement has been driven by radical changes in the quality and distribution of intellectual competencies among the citizens of the world. Over the past several centuries there has been a steady increase in the amount and quality of information expected to be at the command of members of the society. Increasingly, critical literacy, numeracy, and specialized skills and knowledge are becoming the currency of modern living. In the technologically developed nations, concern has arisen around the extent to which citizens of each of these countries are keeping up with these changing demands for intellectual competence. Subject matter referenced or discipline-based criteria have emerged as the universal indicators of such competence. One of the related issues has to do with whether standards should be referenced to subject matter mastery or referenced to derivative intellectual abilities and competencies. Traditional achievement tests and end-of-course assessments were offered as examples of the former, while tests of developed

academic abilities (previously called aptitudes) were offered as examples of the latter.

Attention was called to the need for greater symmetry between standards for student achievement, standards for professional performance, standards for institutional capacity and standards for education assessment. The standards movement has been dominated by the concern for student achievement. Efforts at directing equal attention to standards for opportunity to learn have been resisted in the education policy community. It is possible that opportunity-to-learn standards have not progressed for several reasons, many of which also apply constraints on the standards movement in general, including the following.

- There is tension between the forces that favour centralization and those that favour decentralization (in this case, local options with respect to instructional content and teaching/learning processes).
- It may be considered cheaper to emphasize outcome standards than to improve the quality of education services. There is a difference between the government setting goals and standards and the government assuming responsibility for the achievement of high academic standards.
- There continues to be woefully low levels of agreement on the 'how to' in pedagogy. Agreement on the details of 'correct practice' has been difficult to achieve. Adaptation to group, individual and situational differences, in addition to local preference, especially in the United States, constrains agreement on universal practices.
- There is some weak evidence that suggests that the imposition of standards is associated with gains in academic achievement; however, the contributions of the standards movement and its data have not as yet been powerful influences on education policy.

There has emerged a division of labour within the standards movement, which places the responsibility for determining

education goals at the national level. Under this implicit arrangement in the United States, several states have focused on subject matter content and professional development, while opportunity to learn has emerged as a local concern. It was generally agreed that standards have emerged as an instrument for improving education and education achievement in response to cultural and economic globalization, with its pressure to advance world-class competitive positions. In the middle of this movement, issues concerning equity have been introduced both in support of and in opposition to the confluence of education assessment and education standards in the policy uses of student achievement data.

Tensions between discipline-based and competence-based emphases in education assessment are beginning to find some resolution in approaches to performance-based assessment. As elaborated by McGaw, there are several extant conceptions of performance-based assessment; however, 'to call all assessment other than multiple-choice tests performance assessment is to adopt too broad a definition, while to restrict the definition to cover only real-world, on-the-job performance is too narrow'. The conception of performance assessment in the round-table discussion embraced approaches to measurement that included constructed responses, in real-world contexts and usually involving action and sustained effort on the part of the respondent. There was general agreement that performance assessment allows for:

- improved representation of a broad range of developed academic abilities;
- sustained contextualized performance over time in the production of a product; and
- more authentic (true-to-life-experience) expressions of what students know and know how to do.

It was agreed that the issue is not so much one of the legitimacy of the performance as the problem of valid and reliable ways to measure the performance.

Issue 3: Large-scale assessment as policy research

Validity and reliability of measures of student performances also surfaced as an issue in the discussion of large-scale assessment as policy research. In the context of a discussion of international comparative assessments, it was noted that differences in curricula and contexts confront the assessment community with challenges referable to exactly what will be assessed. In the IEA studies of basic school subjects, the test grid for measuring education outcomes was developed based on an analysis of the curricula of the participating countries. These studies also included instruments to measure school and classroom process variables, as well as teacher and student background variables. To compliment these studies, IEA also conducts studies that are not curriculum-based. In the first group of studies, attention is given to the relationship between what is taught, under what conditions and academic achievement. In the second instance, the focus is more on what is learned and, by implication, on what is taught.

IEA recognizes two purposes for international comparative achievement studies: to provide policy-makers and education practitioners with information about the quality of their education in relation to relevant reference groups; and to assist in understanding the reasons for observed differences between education systems. According to Plomp, IEA was founded as a research co-operative with primarily an academic research focus. During the past fifteen years, IEA has begun to focus more on the interests of policy-makers without losing its interest in explanatory research. The design of TIMSS and the uses of its data provide an example of this bifocal approach. Although not every study should have a size and a design as comprehensive as TIMSS, IEA has the strong belief that the conceptualization and design of its studies allow for analyses that meet

the needs of both policy-makers and education practitioners. Thus, the IEA research studies serve several functions, including description, benchmarking, monitoring of quality, international comparisons and understanding reasons for observed differences.

Several issues were identified in the discussion of large-scale assessment as policy research. It was clearly agreed that large-scale assessments should serve to inform our understanding of the status of education development and the processes by which it is achieved, and to inform policy and practice of ways and means by which higher and more universal levels of achievement are obtained. We as conferees agreed that using large-scale assessment data to inform policy and practice would require:

- much more nuanced and systematic data analysis and reporting than is currently the case;
- examination of a wider range of content, for example, 'intellective competence' as the meta expression of the developed ability to apply affective, cognitive and situative mental processes to the solution of problems;
- greater integration of assessment with teaching and learning, and between different kinds of education assessment;
- increased involvement of and communication with the 200 or so different countries of the world, which vary greatly in their approaches to education as well as in the values held with respect to education and academic achievement;
- greater efficiency in data collection and the frequency with which tests are administered;
- better articulation between large-scale assessment research, on one hand, and the processes of, and participants in, teaching and learning transactions; and
- more consideration to the possibility of shifting the emphasis in large-scale international studies from its focus on the tendency to determine national competitive positions, to comparisons that reference

pedagogically relevant policies and practices and that identify common as well as idiosyncratic programme characteristics.

It was unanimously agreed by the participants that the round table afforded a valu-

able opportunity for dialogue on critical issues concerning the role of measurement and evaluation in education policy. The hope was expressed that the round table would stimulate further discussion, research and development.



