IZA04877

Gathering evidence for learning

John Izard
School of Education
RMIT University
<john.izard@rmit.edu.au>

How do we decide whether students have learned? Learning involves changes in knowledge, skills and the sophistication of the strategies employed by the learners. How do we gather the relevant evidence of learning? To measure these changes we need at least two relevant measures, one prior to a particular stage of learning and a later assessment documenting a higher level of achievement. This paper looks at the requirements for using such assessment for teaching and learning and addresses threats to the validity of assessments to evaluate progress.

Deciding whether students have learned is not as simple as it appears. In the fourth edition of *Theories of Learning* Hilgard and Bower (1975) recognise the difficulty of providing an operational definition of learning but suggest that the following is serviceable:

> Learning refers to the change in a subject's behavior to a given situation brought about by his (sic) repeated experiences in that situation, provided that the behaviour change cannot be explained on the basis of native response tendencies, maturation, or temporary states of the subject (e.g., fatigue, drugs, etc.). Hilgard and Bower (1975, p. 17)

(Hilgard and Bower's definition shows little change from their previous edition of *Theories of Learning* in 1962.) They claim that this definition allows an inference regarding "learning" only when there is no other explanation. They specifically exclude such changes due to response tendencies like reflexes (e.g. contraction of one's pupil in changing light), to maturation, fatigue and habituation, and to performance changes due to drugs, intoxicants or illness.

Hilgard and Bower (1975, p. 22) raise six questions about types of learning that are met in everyday life.

1. What are the limits of learning?
2. What is the role of practice in learning?
3. How important are drives and incentives, rewards and punishments?
4. What is the place of understanding and insight?
5. Does learning one thing help you learn something else?
6. What happens when we remember and when we forget?

We might add a few more:

7. Can learning be inferred from performance?
8. Can absence of learning be inferred from lack of performance?

Gathering the relevant evidence of learning

Gathering *relevant* evidence of learning can be the first stage of evaluating learning and taking action as a consequence. The distinction between the evidence-gathering or measurement and the placing of value on the results is an important one. An evaluation of a parliamentary speech illustrates the distinction.

> "… much of what you said was good, and some of what you said was new. But what was good was not new, and what was new was not good." (quoted by Ahmann and Glock, 1971, p.1)

A key issue is that of *relevance*. Radford (1969) addressed this issue in his summing up at an Australian College of Education annual conference on Educational Measurement and Assessment. He drew attention to the perspectives of stakeholders with this story about a review of *Lady Chatterley's Lover*. Although the book had been before the courts in London as a corrupter of morals, the reviewer of the magazine *Field and Stream* had a more favorable view. The review, in part, said:

> This fictional account of the day-to-day life of an English gamekeeper is of considerable interest to outdoor-minded readers, as it contains passages on pheasant raising, on apprehending of poachers, ways to control vermin and other chores and duties of the professional gamekeeper. Unfortunately one is obliged to wade through many pages of extraneous material in order to discover and savour those sidelights on the management of a Midland shooting estate. This book cannot, however, take the place of J.R. Miller's *Practical Gamekeeping*. (quoted from 'A model of the Innovating Process' in Technology, Transfer and Innovation, National Science Foundation, US Government Printing Office, 1967, p.11)

Relevance

The issue of relevance has to take account of intentions. These intentions may concern individual pupils or may be limited to groups (classrooms, schools, regions, systems, or nations). For example, Pang (2002) used a curriculum evaluation model based on work by Stake (1967) and Stufflebeam's context, input, process, and product (CIPP) model (1974, 1984) to evaluate implementation of the Malaysian Smart Schools science curriculum in a large school. Similarly, Chantachak (2003) investigated the acceptability of new technical language about measurement, assessment and examinations in the context of an aid program for the Ministry of Education and in Laos. In both cases the evidence collected had to be relevant to the issues being investigated.

What indicators of learning are there?

Traditionally, examinations have been used to assess achievement. Examinations have been recognised as a powerful influence on what happens in schools by tending to dictate what is taught and not taught. Examinations often assess what is easy to measure rather than what is important to measure. A prime example is the exclusive use of multiple-choice test items to assess educational achievement. Although there are considerable cost benefits in using so-called objective assessments (the questions are

selected subjectively) some key issues of relevance arise. Multiple-choice tests assess the ability to *recognise* a correct response from a limited list of given alternatives. Such tests are unable to assess the student's ability to *construct* a response. Since the real world involves instances where there are both constructive and recognition responses required, testing only one of these ignores an important area of human discourse. Since examinations implicitly define what knowledge is valued and what is regarded as not important using only multiple-choice examinations is implying that constructive behaviours are less important and that students will rarely need to construct a response.

Assessment tasks do not have to be written tasks. Assessment tasks can be open-ended tasks (such as writing an essay, or solving a mathematics problem), collections of student work, performances (such as oral presentations and musical acts like playing the piano, swimming, practical tasks in science, and folk dancing), projects and investigations, and products of student work (for example, cooking a cake, creating a working model, or making an article out of wood). Rating scales can be used to judge performances provided that those making the judgments have sufficient expertise and are reasonably consistent with each other. Examiners look for different features, give those features differing emphasis, have different views of complexity and therefore assign differing values to the work and they vary in their assigned marks. They need a rationale for sound assessment.

Representativeness

Many assessments are not representative of the curriculum intentions. This distorts findings about the current state of knowledge of the groups being assessed. This is unfair to the students and their parents. Some of the effects are quite subtle.

- When students are attempting external examinations, the teaching is directed often to what the teacher believes will be on the examination rather than to the curriculum that the examination assesses. If the teacher is successful in prediction, students will obtain a better mark than they deserve in terms of their coverage of the curriculum (and the teacher will attract more students). If the teacher is unsuccessful in prediction, the ground will probably shift – the student will be blamed for not working hard enough.
- Many public examinations, commercially published tests and teacher-made tests favour some students over others. If you are at the level at which most of the test items are pitched you will have many opportunities to find an item within your capability. But if you are more able or less able there will be fewer items that are pitched at your level. This discrimination is rarely acknowledged, although it has a direct bearing on the usefulness of the scores for selection for further study or employment. For example, universities complain that they cannot make use of examination scores to distinguish between the best students for the purpose of awarding scholarships. If we assume that such examination scores should be available for these purposes, the students have to engage with challenging questions at the highest level. Unless students face questions that they cannot answer we do not know what they do not know. A second example is the claim that students lack skills for further study or employment. Who is at fault? Is it the university or employer that uses an inappropriate selection device based on rank order rather than desired skills? Or is it the examining board that is confined by cost and political expectations to make assessments on an extremely small sample of student work?

Inappropriate use of evidence

When the reports of the first national *Australian Studies of School Performance* (Keeves and Bourke, 1976; Bourke and Lewis, 1976) were published in 1976 there were many claims in the media that standards had fallen. Journalists did not appear to understand that at least two measures are required to identify whether a change has occurred. (When the corresponding study some five years later reported that standards had not fallen, and in some areas had improved, no newspaper in Australia reported the findings.)

As stated earlier in this paper, examinations have been used to assess achievement. Although examinations usually only show a current state of knowledge, the results are often interpreted as inferring that learning has occurred. This inappropriate conclusion is a consequence of lack of understanding about research methodology. Without knowing the prior state of knowledge we cannot know whether learning has occurred or whether the teaching was successful. (Izard, 2004)

There are several ramifications of this lack of understanding. At the disadvantaged end of the educational spectrum, those seeking to fund educational intervention fail to evaluate or acknowledge the current stage of learning before setting targets. When inappropriate targets are set for educational interventions some schools appear ineffective because the target is impossibly high, while other schools appear very effective because they had surpassed the target before it was set. (This may be a consequence of employing "managers" who know little about what they are managing and do not accept that they need to know.) The reports of such interventions are of little use in determining "what works" because of the faulty methodology.

At the more privileged end of the educational spectrum, the refusal to look at the value-added component of education results in schools being credited with sound teaching without the contribution of their selective entry policies being considered. Such schools are described as "good" whether they deserve the label or not. This then leads to the simplistic use of averages by politicians to stigmatize half of the population (without realising that half of any Federal Cabinet is below the Cabinet average – by definition – in competence or honesty or integrity or whatever other variable is chosen). [Perhaps the extra funding offered to such schools is a reward for their elitist selection approach that implies that only some are worthy of an education. This is similar to the view that only politicians are guaranteed their retirement perks in full. The rest of the population has to suffer the reductions imposed by high management costs unrelated to effort or benefits, economic recession, and in some cases the theft of compulsory deductions from worker wages and salaries as a consequence of ATO failure to ensure receipt of contributions.]

Failures to gather relevant evidence

To measure changes we need at least two relevant measures, one prior to a particular stage of learning and a later assessment documenting a higher level of achievement. But measures that are apparently appropriate may still mislead if there is:

- inappropriate distribution of item difficulty (providing selective evidence),
- failure to acknowledge ceiling and floor effects when making multiple assessments (hiding effects of intervention), and

- restriction of numbers of participants in the investigation of educational interventions (limiting the possibility of detecting improvement),

The first two of these issues will be taken up again below in the discussion about validity.

Reporting the evidence in non-useful ways

One of the most serious problems with teacher-made assessments and many external examinations is that the data are reported in ways that militate against the information gathered being used to improve teaching and learning. A test score is interpreted with respect to the *group*, not to the *items*. Performance by an individual is compared with reference group scores and expressed as a proportion of *students* (not *items*). The items that were used to obtain these scores are ignored in the interpretation. [Often the reference groups are not relevant for the comparisons that are being made. The use of "reading ages" is an example of inappropriate practice.]

The traditional ways of reporting scores are *not* in terms of how well the student has satisfied each component of the curriculum. Those interpreting traditional test results appear to assume that doing better than "average" is good. (They ignore the fact that, logically, half of the group is better than "average".) Achievement is reported in terms of place in class or ranked categories (often confused further with meaningless letter grades) or in terms of place in a cohort. Such reports, typical of *assessment **of** learning* are little better than rank orders and mask any evidence that teaching/learning is improving or getting worse (Izard, 2004). When teachers do look at individual performance on items, some make the mistake of teaching students how to do particular test items rather than teaching the skills that were tested by that item. Since subsequent tests may use different items, teaching the test item rather that the relevant skills appears to give students skills but performance on the taught items does not necessarily transfer to performance on similar items.

By contrast, the approach known as *assessment **for** learning*, (sometimes known as *formative assessment*) describes what each student already knows. Assessment information should inform the teacher (and the student) of what tasks can be attempted successfully, what skills and knowledge are being established currently, and what skills and knowledge are not yet within reach. When teachers have this information, they focus on what students need to learn, and students show evidence of learning. (Izard and Jeffery, 2003)

How do we check that the indicators of learning are valid for their purposes?

The quality of the assessment strategy for evaluating progress is compromised if there are insufficient relevant items to show curriculum effectiveness, if inappropriate statistics are used for reporting, if valid measures of change are lacking, and if there is a shortage of assessment expertise to interpret the evidence. (More-of-the-same testing may not be a solution since selective evidence is not generally representative and additional time devoted to *testing* detracts from time for *teaching* and may duplicate effort.)

The need for valid indicators of learning can be illustrated by interpreting a diagram from a paper presented earlier this year. (Izard, 2004)

	Test A	Test B	Test C	Test D	Test E	Test F	Test G	Test H	Test I	Test J
		B5				F5	G5		I5	
		B4				F4	G4		I4	
		B3				F3				J5
		B2								
		B1								
x1										
			C5							
			C4						I3	
			C3						I2	J4
			C2							
			C1							
x2										
				D5						
				D4			G3			
				D3			G2		I1	J3
				D2				H5		
				D1						
x3										
					E1					
					E2			H4		
					E3			H3		J2
					E4			H2		
					E5					
x4										
	A5									
	A4									
	A3					F2				J1
	A2					F1	G1	H1		
	A1									

Figure 1 Alternative possibilities for tests (Source: Figure 6 in Izard, 2004)

In the diagram the achievement continuum is represented by the broken vertical line within the frame. Achievement levels of four students (as established from extensive testing) are shown at the left as x1, x2, x3 and x4. Student x1 has the highest achievement and student x4 has the lowest achievement. The diagram could also

represent four sets of results for the same student with x4 representing the initial achievement, x3 after some learning has occurred, x2 after further learning, and x1 at the final stage.

But the properties of the indicators of learning (tests) may not show differences in achievement. In the simplest cases, Test A with easy questions could not distinguish between the four achievement levels. After a period of learning, this test would not be able to detect improvements in achievement because one cannot obtain any higher than a perfect score. (This is known as a ceiling effect.) Such lack of improvement may be interpreted as a failure to learn but the fault lies with the construction of Test A. (In an intervention context, use of minimum competence tests over a narrow range of achievement has a similar effect of masking the learning that occurs.) Similarly, Test B with difficult questions would not distinguish between the four achievement levels. (This is known as a floor effect. For example, questions appropriate for Grade 7, Grade 12, 1st Year undergraduate, and post-graduate levels could not usually distinguish between the consequences of learning at pre-school and the first three years of school.) Lack of improvement may be interpreted as a failure to learn but the fault lies with the construction of Test B.

Tests C, D and E separately do not distinguish between all the achievement levels. Test F combines the easy questions of Test A and the difficult questions of Test B but is still not able to distinguish between the achievement levels. Results on Tests A, B and F would imply that the 4 achievement levels were the same even though we know from other evidence that they are different. Tests G, H, and I distinguish between some achievement levels but the scores would not reflect the actual differences between the achievement levels. Tests G, H and I do not have an even distribution of item difficulty for the range of achievement. These tests are discriminatory. The tests differ in providing an opportunity to obtain a higher score.

Test J avoids this discrimination by using a rectangular distribution of item difficulty. Test J would be useful in distinguishing between each of the achievement levels. Further, the differences between the attainment levels are reflected in the differences between the scores. Test J is the only test design that can detect learning changes over time where the difference in scores may be interpreted as the extent of the success of the learning.

Although this example was based on open-ended items scored right or wrong, the same ideas apply to partial credit items. (Partial credit items have more than one mark available. For example, one partial credit item may receive a score of 0, 1, or 2 while another may receive a score of 0, 1, 2, or 3, and so on.) These ideas also apply to multiple-choice items where the test (or sub-test) is long enough for the influences of random guessing to have reduced effect on achievement scores. For the purposes of explanation this discussion has been confined to small tests. In practice the tests should be much longer (but retain the rectangular item difficulty distribution) in order to make valid inferences about achievement and improved learning.

The key validity issue in using *assessment **for** learning* (or testing for teaching purposes) is *the extent to which meaningful, appropriate and useful inferences can be made about scores*, and many tests used to assess achievement lack the essential qualities to provide these meaningful, appropriate and useful inferences about scores.

A further difficulty with teacher-made (and lecturer-made) test items is that many such tests are conceived and written without a specification to ensure adequate sampling of topics and sub-tests often have too few items to provide meaningful information. Test items constructed by teachers without technical support do not always distinguish between knowledgable students and less knowledgable students. Respectable commercial test publishers do check that items are useful in this way, but most teachers and lecturers lack the skills and the time to carry out the analyses. Tests differ in number and format of items. Items vary in difficulty but are treated as equivalent in difficulty. Two tests of the same topic are assumed to cover the same work without empirical validation (Izard, 2004). Izard (1998a, 2002a) reviews other constraints in giving candidates due credit for their work in papers on strategies for quality control in assessment.

What are appropriate statistical strategies for analysing evidence of learning?

If two tests are given at differing times (without being given together to this group or another comparable group) any changes in the score cannot be interpreted. *One does not know whether the difference in scores was because the tests differed in difficulty, whether learning occurred over the time interval, or whether some combination of these events occurred*. Many teachers do not know the relative difficulties of the tests they give, so interpreting the results from their tests is impossible if they have not used the same test twice. But that would encourage teaching to the test instead of to the curriculum so that is not appropriate either.

Both schools and tertiary courses face this problem of the lack of information on the progress made by students over several year levels. This is a consequence of using different tests at different stages of learning without ever asking how the scores on each test relate to the overall continuum of achievement in that subject or learning area. For example, how does the difficulty of the first year undergraduate examination compare with the difficulty of the second and third year undergraduate examinations in the same area of study.

Research at tertiary undergraduate level in Britain (Izard, *et. al*., 2003) addressed this issue by scaling the results to adjust for tests that differed in difficulty. Differences in scaled scores were found because learning occurred (or skills were forgotten) over the time interval.

What do we understand about consequent actions to be taken after considering the evidence of learning?

Improving achievement requires more than good assessment. To use a farming analogy: the farmer's maize will grow better if appropriate nutrients and water are provided in timely fashion. *Measuring the height of the maize frequently is not going to improve the yield at all*. To provide effective learning opportunities for all pupils we need assessment strategies that will be teacher-friendly (helpful in identifying what has to be taught) and teaching strategies to ensure students learn what they currently do not know (Izard, 1998b, 2004).

Using assessment for learning has implications for the design of the assessment instruments and the way the information gathered with those instruments is reported. Since the assessments have to be representative of the curriculum intentions (not just the pencil-and-paper components), the assessment specification has to include each of those intentions. Secondly, assessment for learning implies that there will be more than one comparable assessment otherwise progress will not be revealed. Practical considerations about undesirable teaching to the test (instead of to the curriculum) imply that the subsequent tests should be different from earlier tests.

Achieving comparability between tests that are given at different times presents difficulties. Assessment for learning requires tests that are given at different times to the *same* students. The meaning of comparability is different in this context. It would be inefficient to give two *parallel* tests. At the time of giving the first test many of the items could be considered too difficult because the concepts had not been taught and, if the teaching/learning process had been successful, at the time of giving the second test many of the items could be considered too easy. Students would find the earlier test daunting and may be discouraged from engaging in the learning that the curriculum intends. The development of suitable instruments for assessing the progress of learning requires an explicit domain or continuum and the tests to provide evidence of learning have to sample the stages in the continuum.

Conclusion

Appropriate assessments to inform students and teachers about what students know and do not know are available. The methods of analysis are well established (see for example, Wright and Stone, 1979; Wright and Masters, 1982; Wilson, 1992). These approaches have been used in longitudinal studies about attainment of statistical ideas (for example, see Watson, Kelly and Izard, 2004 [AARE search code WAT04867]), in studies about science learning and conceptual understanding in social education (for example, see Adams, Doig, and Rosier, 1991; Doig, *et al*., 1994), and in national and international surveys of attainment (for example, see the *Trends in International Mathematics and Science Study* [TIMSS, formerly known as the *Third International Mathematics and Science Study*] at http://www.nces.ed.gov/timss/, and the Organisation for Economic Co-operation and Development [OECD] Programme for International Student Assessment [PISA] at http://www.pisa.oecd.org/). But much assessment effort is wasted if the information is not **used**. It seems obvious that one should seek to teach students what they do not know, rather than continue to teach them what they already know and have demonstrated. These approaches have been used at school level to inform teachers about needed learning and to monitor individual progress over time (for example, see Izard, 1998b; 2002b, 2002c ; Izard, Jeffery, Silis, and Yates, 1999; Izard and Jeffery, 2003).

Those using or developing instruments for assessment of learning must address several important issues. Firstly, they need to ensure that their assessment tasks represent the relevant domain or continuum. Secondly, they need to ensure that those tasks will provide evidence of the progress the student has made along the continuum. Thirdly, they need to ensure that effective teaching occurs (Black and Wiliam, 1998a, 1998b). This may well reduce predictive validity: students not expected to be successful turn out to be successful when good use is made of formative assessment (Izard, 1980). Without achieving success with respect to these issues, they will not be able to address the issue

of assessing each student's learning in a sound way. Nor will they be able to gauge the effectiveness of their teaching.

References

Adams, R.J., Doig, B.A. & Rosier, M.J. (1991). *Science learning in Victorian schools: 1990*. (ACER Research Monograph No. 41). Hawthorn, Vic.: Australian Council for Educational Research.

Ahmann, J.S. and Glock, M.D. (1971). *Evaluating Pupil Growth: Principles of Tests and Measurements*. (3rd ed.). Boston: Allyn and Bacon.

Bourke, S.F. and Lewis, R. (1976) Literacy and Numeracy in Australian Schools: Item Report. Australian Studies in School Performance. Volume II. Canberra, ACT.: Australian Government Publishing Service.

Black, P. and Wiliam, D. (1998a) "Assessment and Classroom Learning," *Assessment in Education, Vol. 5*, pp. 7-74.

Black, P. and Wiliam, D. (1998b) "Inside the Black Box: Raising Standards Through Classroom Assessment," *Phi Delta Kappan*, p.139 (Also at www.pdkintl.org/kappan/kbla9819.htm)

Chantachak, P. (2003). *Difficulties in interpreting educational concepts from English to Lao language in ASLO (Assessment of Students' Learning Outcomes) workshops*. Unpublished masters thesis, RMIT University, Melbourne, Australia.

Doig, B.A., Piper, K., Mellor, S. & Masters, G. (1994). *Conceptual understanding in social education*. (ACER Research Monograph No. 45). Melbourne, Vic.: Australian Council for Educational Research.

Hilgard, E.R. and Bower, G.H. (1975). Theories of Learning (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Keeves, J.P. and Bourke, S.F. (1976) Literacy and Numeracy in Australian Schools: A First Report. Australian Studies in School Performance. Volume I. Canberra, ACT.: Australian Government Publishing Service.

Izard, J.F. (1980). *An investigation of the effects of spatial and other abilities on children's performance in area, volume, and related aspects of school mathematics curriculum*. . Unpublished doctoral thesis, La Trobe University, Bundoora, Australia.

Izard, J.F. (1998a). Quality assurance in educational testing. In National Education Examinations Authority (Eds.) *The effects of large-scale testing and related problems: Proceedings of the 22nd Annual Conference of the International Association for Educational Assessment*. (pp.17-23). Beijing, China: Foreign Language Teaching and Research Press.

Izard, J.F. (1998b). Validating teacher-friendly (and student-friendly) assessment approaches. In D. Greaves & P. Jeffery (Eds.) *Strategies for intervention with special needs students*. (pp.101-115). Melbourne, Vic.: Australian Resource Educators' Association Inc..

Izard, J.F. (2002a). Constraints in giving candidates due credit for their work: Strategies for quality control in assessment. In F. Ventura & G. Grima (Eds.) *Contemporary Issues in Educational Assessment*. (pp. 15-28). MSIDA MSD 06, Malta: MATSEC Examinations Board, University of Malta for the Association of Commonwealth Examinations and Accreditation Bodies.

Izard, J.F. (2002b). Describing student achievement in teacher-friendly ways: Implications for formative and summative assessment. In F. Ventura & G. Grima (Eds.) *Contemporary Issues in Educational Assessment*. (pp. 241-252). MSIDA MSD 06, Malta: MATSEC Examinations Board, University of Malta for the Association of Commonwealth Examinations and Accreditation Bodies.

Izard, J.F. (2002c). Using Assessment Strategies to Inform Student Learning. In P. Jeffery (Compiler)*: Proceedings of the Annual Conference of the Australian Association for Research in Education Brisbane December 2002*. (http://www.aare.edu.au [search code IZA02378]). Melbourne: Australian Association for Research in Education.

Izard, J.F. (2004). Best practice in assessment for learning. Paper presented at the Third Conference of the Association of Commonwealth Examinations and Accreditation Bodies on *Redefining the Roles of Educational Assessment*, March 8-12, 2004, Nadi, Fiji: South Pacific Board for Educational Assessment.

Izard, J.F., Haines, C.R., Crouch, R., Houston, S.K., and Neill, N. (2003). Assessing the impact of the teaching of modelling: Some implications. In S.J. Lamon, W.A. Parker, and K. Houston (Eds.) *Mathematical Modelling: A Way of Life: ICTMA 11*, (pp. 165-177.) Chichester: Horwood Publishing

Izard, J. and Jeffery, P. (2003). Testing for Teaching: A longitudinal formative assessment project. Paper presented at the joint NZARE-AARE Conference in Auckland, November-December 2003. (http://www.aare.edu.au [search code JEF03075]). Melbourne: Australian Association for Research in Education.

Izard. J., Jeffery. P., Silis, G.F., and Yates, R. L. (1999). Testing for Teaching Purposes: Application of Item Response Modelling (IRM) teaching-focussed assessment practices and the elimination of learning failure in schools. In Peter Westwood & Wendy Scott. (Eds.) *Learning Disabilities: Advocacy and Action* (p 163-188). Melbourne. Australian Resource Educators' Association Inc. (AREA)

Pang, V. (2002). *The Malaysian Smart School Curriculum: An Evaluation Case Study of an Implementation*. Unpublished doctoral dissertation, RMIT University, Melbourne, Australia.

Radford, W.C. (1969). Some changing attitudes to and views on evaluation and measurement. In *Educational Measurement and Assessment*. (pp. 149-163). Carlton, Vic.: Australian College of Education

Stake, R. E. (1967). The countenance of educational evaluation. *Teachers College Record, 68*, 523-540.

Stufflebeam, D. L. (1974). Alternative approaches to educational evaluation: A self-study guide for educators. In W. J. Popham (Ed.), *Evaluation in Education* (pp. 97-143). Berkeley, CA: McCutchan Publishing Corp.

Stufflebeam, D. L. (1984, September 2-5). *Improvement oriented evaluation*. Paper presented at the Evaluation of Educational and Training Programs Workshop, Canberra, Australia.

Watson, J., Kelly, B. and Izard, J. (2004). Student change in understanding of statistical variation after instruction and after two years: An application of Rasch analysis. Paper presented at the AARE Conference in Melbourne, Nov.-Dec. 2004. (http://www.aare.edu.au [search code WAT04867]). Melbourne, Vic.: Australian Association for Research in Education.

Wilson, M. (1992). Measurement models for new forms of assessment. In M. Stephens & J. Izard. (Eds.) *Reshaping assessment practices: Assessment in the mathematical sciences under challenge*. (pp. 77-98). Melbourne, Vic.: Australian Council for Educational Research.

Wright, B.D. & Masters, G.N. (1982). *Rating scale analysis*. Chicago, IL.: MESA Press.

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