Pencil-and-paper Mathematics Tests Under the Microscope

Nerida F. Ellerton *and* **M. A. ("Ken") Clements**
Faculty of Education **Faculty of Education**
Edith Cowan University **The University of Newcastle**
Churchlands, Western Australia 6018 **Callaghan, NSW 2308**

The effectiveness of a highly regarded short-answer pencil-and-paper test *(Mathematics Competency Test,* published in 1996 by the Australian Council for Educational Research) was investigated. In the study, 182 students—in Years 5 through 8, in 8 classes in NSW and Western Australia—answered the test questions and were interviewed. About 28% of responses were *either* (a) correct but students showed less than full understanding; *or* (b) incorrect, but students showed at least partial understanding, of the key concept/s) and skill(s).

According to *Measuring Counts,* a policy brief issued by the Mathematical Sciences Education Board (1993) in the United States, pencil-and-paper tests became popular because they provided an efficient means of assessing large numbers of people at low cost. There were, however, major criticisms of the tests: they could not tap students' ability to estimate the answer to arithmetic calculations, to construct geometric figures, to use calculators or rulers, or to produce complex, deductive arguments. Furthermore, the psychometricians' narrow focus on technical criteria—primarily reliability—worked against educationally sound assessment because it generated tests comprising many short items, and did not permit examinations composed of a small number of complex problems. This resulted in students being asked to perform large numbers of smaller tasks, "each eliciting information on one facet of their understanding, rather than to engage in complex problem solving or modelling" (p. 7).

In the 1990s there has been an awareness of the deficiencies of externally-set, pencil-and-paper mathematics tests (hereafter denoted ESPPMTs). Kilpatrick (1993), for example, called for less emphasis to be given to the psychometric aspects of assessment. According to Kilpatrick (1993), educators need to understand how people, "not apart from but embedded in their cultures, come to use mathematics in different social settings and how we can create a mathematics instruction that helps them use it better, more rewardingly, and more responsibly" (p. 44). To do that, he added, "will require us to transcend the crippling visions of mind as a hierarchy, school as a machine, and assessment as engineering" (p. 44).

Rhetoric and Reality

The Rhetoric

Over the past 10 years, in fact, much has been said, verbally and in writing, by advocates of more authentic methods of assessing mathematics learning and teaching (Black, 1994; Clarke, 1988, 1996; Clements & Ellerton, 1996; de Lange, 1996; Holt, 1993; Kilpatrick, 1993; Leder, 1992; Mathematics Sciences Education Board, 1993; National Council of Teachers of Mathematics, 1995; Niss, 1993; Romberg, 1993; Webb & Coxford, 1993). In the United States of America, the National Council of Teachers of Mathematics (NCTM) not only emphasised the need to develop and introduce new assessment methods in its *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989), but it also published, in 1995, a special volume on *Assessment Standards for School Mathematics* (NCTM, 1995).

The Reality

Despite the rhetoric calling for more authentic assessment procedures to replace, or at least complement, ESPPMTs, mathematics educators and education bureaucrats have found it difficult to escape from a psychometric mindset in which the concepts of test reliability, validity, item discrimination, etc., are set in stone. Strident public

demand for education accountability has given rise to rigid interpretations of "best-practice" education environments in which ESPPMTs have an important role to play. Best practice, many bureaucrats believe, can be most effectively achieved by specifying outcomes and behaviours indicating that these outcomes have been achieved, and then asking test experts to construct high-quality, highly reliable and valid, pencil-and-paper tests which will not only monitor the extent to which the outcomes are being achieved in different schools, but also (through the process of benchmarking) enable schools to be ranked according to "value-added" criteria (Cuttance, 1993; McGaw, 1995).

Teachers, education bureaucrats, politicians seem to be willing to allow *formative* assessments to be carried out by "alternative modes" (Clarke, 1988, 1996; Schmidt & Brosnan, 1996), but for *summative* tests, the demands of public accountability ostensibly require valid and reliable pencil-and-paper instruments, created by test experts, to be used (Black, 1994).

It is one thing for educators to criticise current practice and to propose new approaches, but it is altogether another thing to achieve large-scale change. Late in the 1990s ESPPMTs still continue to be widely used. Indeed, throughout the 1990s, education bureaucracies throughout the Western world, including Australia, New Zealand, and the United Kingdom, have been increasingly driven by societal demands for greater school and system accountability (Australian Education Union, 1995; Black, 1994; Garet & Mills, 1995; Olssen, 1993).

The move towards accountability led Shavelson, Baxter, and Pine (1992) to comment that in the final analysis the United States of America might be placing far too much weight on accountability to achieve its reform agenda. They argued that states with the strongest and most technically sound accountability systems had not achieved their desired reforms, and stated:

> Perhaps what is needed is far less account taking and far greater consideration and resources given to teaching and learning, especially for students drawn from diverse social, economic, cultural and language backgrounds. An increased emphasis on and bully-pulpit use of highstakes testing may, paradoxically, have a deleterious effect on U.S. education. Tactical, political "fixes" are not what is needed. Rather, we believe that a long-term, realistic approach to assessment—one that transcends politicians' terms of office—is. (pp. 26–27)

Political and professional realities render it almost certain that ESPPMTs will continue to be widely used across the world for assessing the mathematics performance of school children.

Yet, it is difficult to mount serious public arguments against ESPPMTs when the largest mathematics competition in the world, the Australian Mathematics Competition, uses such tests in our own backyard. So does the International Association for the Evaluation of Educational Achievement (IEA) which, with funds provided by governments around the world, has used pencil-and-paper tests in its three major international mathematics achievement studies.

Stake (1995), the veteran North American education evaluator, has recently stated that moves towards greater accountability in schools, with their attendant reliance and pencil-and-paper tests, are dangerous, in that they are likely to generate "overstandardisation, oversimplification, over-reliance on statistics, student boredom, increased numbers of dropouts, a sacrifice of personal understanding, and probably a diminution of diversity in intellectual development" (p. 213).

According to Stake, psychometricians' attempts to define different kinds of test validity are misleading because "testing as an activity and individual tests as a tool are neither valid nor invalid until the results are interpreted in some way," for "only the interpretations of test scores in particular situations can be said to be valid or invalid" (p. 172). Stake went on to say that research indicates that with increased testing and curriculum standardisation, teachers attend more to the so-called basics—the most elementary knowledge and skills—and less to what is needed to get their

students to achieve a deep understanding of even a few topics.

Research Into the Effectiveness of Externally-set, Pencil-and-paper Mathematics Tests of the Short-answer Variety

Despite the continuing assumption by education bureaucrats (and parents, politicians, and many teachers), that externally-set, short-answer pencil-and-paper mathematics tests provide a cost-effective, objective way of monitoring mathematics performance, it needs to be recognised that recent mathematics education research is calling into question the power of externally-set written examinations to provide proper education accountability.

Summary of Research into the Effects of Standardised Testing

Kindsvatter, Wilen and Ishler (1988) drew attention to a substantial body of literature which describes the effects of large-scale, standardised testing for accountability. Data indicate that the use of such tests does not raise standards, and teachers consider that standardised tests do not assist learning. In fact, there is no conclusive evidence to show that assessment of student learning from externally-set pencil-and-paper tests provides better quality data than could be provided by teachers' ratings of students (Mathematical Sciences Education Board, 1993).

Furthermore, it is not widely known that research has generated data suggesting that, despite enthusiastic claims to the contrary by psychometricians and education bureaucrats, performance data generated by ESPPMTs can be seriously misleading. There is, in fact, increasing evidence pointing to the conclusion that students who answer pencil-and-paper mathematics items correctly sometimes have little or no understanding of the mathematical concepts and relationships which the items were designed to measure, and that this applies even for so-called "valid" and "reliable" tests (Frary, 1985; Gays & Thomas, 1993; Hembree, 1987; Thongtawat, 1992).

Nevertheless, politicians and education bureaucrats still believe that ESPPMTs can satisfactory measure mathematical learning. A large research study into this question, jointly supervised by the authors and carried out in Thailand by Thongtawat (1992), found that the proportion of students who gave correct answers to multiple-choice mathematics items but who did not understand the mathematical concepts and relationships involved in the items, was much higher than that for corresponding short-answer but not multiple-choice items. Thongtawat also found that students who scored poorly on a test could sometimes have a good conceptual grasp of the material which the items covered.

The authors have previously carried out research designed to test the effectiveness of multiple-choice and short-answer pencil-and-paper tests (Ellerton & Clements, 1995). In a study in which 116 Year 8 students in 5 schools in 2 Australian states were interviewed, we found that both multiple-choice and short-answer questions were seriously ineffective in assessing student understanding. We concluded:

> We are particularly concerned about the increasingly widespread use of externally-set, multiple-choice and short-answer pencil-and-paper instruments. Testing regimes, based on crude, one-off, pencil-and-paper instruments are employed, and justified by reference to the need to make teachers, schools and systems more "accountable." The data in this paper suggest that claims that such tests can provide "quality assessment data" about students' and schools' mathematical performances are not justified. (pp. 12–13)

In fact, over one-third of responses to well-constructed questions in multiple-choice and short-answer formats were such that either (a) correct answers were given by students who did not have a sound understanding of the correct mathematical knowledge, skills, concepts and relationships which the questions were intended to cover; or (b) incorrect answers were given by students who had partial or full understanding. Almost 50% of the incorrect responses were given by students who

did understand, at least partially, the mathematics that the questions were designed to assess (Ellerton & Clements, 1995).

In the same paper, we questioned whether multiple-choice questions could ever be valid for assessing mathematical learning, pointing out that we do not know of any practising adult mathematicians who actually work in situations where they are regularly asked to choose one correct answer from four or five possibilities. The implication is that schools which use multiple-choice tests as a major method for assessing mathematical learning are in danger of continually reinforcing in their students' minds a flawed image of the nature of mathematics.

The Present Study

The Instrument

In our most recent ("in progress") research, we are investigating the effectiveness of the *Mathematics Competency Test,* published in 1996 by the Australian Council for Educational Research. This 46-item test has an excellent pedigree. Its three authors (P. E. Vernon, K. M. Miller, and J. F. Izard) are highly regarded educationists/psychologists/psychometricians; it is distributed by a highly regarded education research organisation; and the manual for the test indicates it has construct, content-related, and criterion-related validity, and an internal consistency index of 0.94 (Izard & Miller, 1996). The test, which is designed for individuals 11 years-of-age or older, uses short-answer but not multiple-choice items, and therefore the guessing factor is minimised.

The Sample

Thus far, 182 students who took the test have been interviewed by graduate research assistants. The students were in 8 classes in three schools—a middle-class independent school in New South Wales (72 students in four Years 5 and 6 classes were interviewed); a government school in the Western Suburbs of Sydney (47 students in two Year 8 classes); and a highly regarded government school in a working class suburb of Perth (63 students in two Year 7 classes). On average, each interview took about 40 minutes.

The Aim of the Study

The aim of our study was to determine, for each particular item, and for each particular student interviewed, the extent to which the student understood the main concept which that item was intended to test, and then to classify initial responses as being "matches" (when a correct response is associated with an adequate understanding of the key concept(s) being tested, or an incorrect (or no) response is associated with an inadequate understanding) or "mismatches" (when a correct response is associated with inadequate understanding or an incorrect (or no) response with an adequate understanding).

Following the classification of responses, the proportions of responses which were deemed to be "matches" were calculated for (a) each interviewee, (b) each class of students, (c) each school, and (d) the total group of interviewees.

The Extended Newman Interviews

The authors have previously developed and used (Ellerton & Clements, 1995) an extended form of the Newman error analysis technique (Ellerton & Clarkson, 1992; Newman, (1977) to investigate *both the errors and correct answers* given by students to items on pencil-and-paper mathematics tests. In the present study the 182 students were interviewed by 4 graduate assistants who had been trained, by the authors, to conduct extended Newman interviews. Although the students had originally attempted all 46 questions on the *Mathematics Competency Test,* extended Newman interviews were conducted for only 33 of the questions—the first 33 questions for students in Years 5 and 6 and, depending on a student's initial score on the test either Questions 1 to 33 or Questions 14 through 46 for students in Years 7

and 8 (students who obtained high raw scores were interviewed for Questions 14 through 46).

The aim of each interview was to ascertain the level of understanding associated with each of the responses by each student to the 33 questions on the *Mathematics Competency Test.* After the standard Newman questions had been asked, and answers given, the interviewer then carried out probes into issues arising from answers given to the questions and from answers given during the interview.

For each question, a student's "Level of Understanding" (LU) was assessed, by the interviewer, according to a 3-point scale: "0," if a student did not recognise, or had no grasp of the necessary concepts; "1," if a student recognised which concepts might apply but had only a limited understanding of the concepts; and "2," if the concepts and relationships were recognised and were understood. Criteria for making these decisions had been developed before the main study began and, following each set of interviews, the interviewers met with the authors to discuss any difficulties they had experienced in making LU classifications. These meetings resulted in LU classification criteria being clarified and sharpened.

Three examples of LU classifications: Question 20 on the *Mathematics Competency Test* stated: "A block of wood measures 6 cm by 5 cm by 4 cm. What is its volume?" Many of the students gave "120" as their written answer, without stating the unit. According to the "Scoring Key" in the *User's Manual,* this response had to marked as incorrect. However, in the interviews some of the students who gave this answer stated confidently, without any prompting, that the volume was 120 cm^3. These students were given a LU classification of 2 for the question.

Question 22 on the *Mathematics Competency Test* stated: "$3y + 2 = 14$ $y =$ " Most of the students interviewed did not know the convention that "3y" means "3 times y." Some guessed that this was the case, and gave the correct answer (y = 4). Others guessed that $3y$ meant $3 + y$, and gave the answer $y = 9$." When told, in the interview, that $3y$ meant "3 times y," many of these students had no trouble working out that y should be 4. These students had their initial response marked "wrong," but were given an LU classification of 2 (because, it was decided, the main point of the question was whether a respondent could solve the "equation" "If 2 is added to 3 times a number, and the answer is 14, what is the number?")

Question 31 on the *Test* showed three boxes which were labelled A: "200g $4.00" B: 250g $4.50 C: 300g $6.00 D: 500g $9.50, and respondents were asked to state which box—A, B, C or D—represented the best value? Many who selected the correct alternative (B), had either guessed, or, in the interview, provided an inappropriate reason. Often they changed their answer in the interview. Such students were given a LU classification of 0 even though their initial answer was correct.

Results

It is not possible to provide detailed analyses of results for individual questions or students in this paper—these will be given, however, in subsequent papers, and in a final report which will be written after the sample of schools and students involved in the study has been enlarged.

Overall, about 28%—that is to say, over one-quarter—of the original responses were "mismatches," in the sense that correct answers were given by students who lacked an adequate understanding of the key concept(s) being tested, or incorrect answers were given by students with full or partial understanding. No student had no mismatches, and more than half the responses of some students were "mismatches."

The proportion of responses which were mismatches for intact classes of students varied from 20% (for a Year 8 class) to 41% (for a Year 7 class). For the three schools, the percentages of mismatches were 23% (for the independent school), 21% (for the government high school) and 41% (for the government primary school).

Discussion

Although the *Mathematics Competency Test* has a reliability index of 0.94, over one-quarter of the students' initial responses were classified as "mismatches." Our data indicate that raw scores on the test do not permit informed judgments to be made about students' knowledge and understandings with respect to the mathematics covered by the test. A correct answer did not always correspond to "understanding" and, for an educationally significant proportion of responses, students who did understand the main concept(s) involved in a question gave incorrect answers.

The findings of our research fly in the face of continued widespread usage of pencil-and-paper tests to assess mathematical understandings in all parts of the world. There are strong vested interests which are propagating the view that ESPPMTs provide the most objective and cost-efficient means of measuring mathematical learning. In Australia, full cohort statewide testing is already taking place, and national testing is on the Federal Government's agenda. Yet, our evidence would suggest that if pencil-and-paper mathematics tests are being used (as is indeed the case) then it is inevitable that invalid results will be obtained.

Validity and reliability issues: Research into why elementary and junior secondary school students make mistakes on word problems on pencil-and-paper tests has consistently indicated that well over half of the errors are associated with difficulties with what Newman (1977) called Reading, Comprehension, or Transformation (Ellerton & Clements, 1992). Analysis of written responses, without any other form of analysis, does not reveal the proportion of correct answers given by students who have little or no understanding of the mathematics involved, or the proportion of incorrect answers which were given by students with good understanding.

This raises serious questions about whether ESPPMTs can ever generate "valid" assessments of mathematical understandings (especially for tests designed for primary and junior secondary students). At issue, in fact, is the meaning of the term "valid," given that the interviews revealed a large number of personal and linguistic variables which affected students' responses. Such variables are usually not recognised fully by test constructors who rely heavily on traditional psychometric considerations (such as within-test correlations—Izard & Miller, 1996, p. 28).

Izard and Miller (1996) recognised some of the limitations of their test, so far as validity was concerned. They stated, in the *User's Manual:*

> Although the test items are consistent with those used in similar instruments, they do not include some types of mathematical task. This is a function of the mode of testing; all other mathematics tests like this suffer the same deficiency with respect to construct validity. The authors of this test recognise that the mode of assessment with a written test provides little evidence of a candidate's skills in solving complex problems in context, undertaking investigations, or carrying out particular practical mathematical tasks such as estimating distances in a local context, measuring using a variety of units, and manipulating space. Accordingly, a comprehensive assessment of mathematical competence will need to combine evidence of performance on traditional written test questions (such as those on the Mathematics Competency Test with judgments about the candidate's proficiency in these other tasks. (pp. 25–26)

This quotation is important, for it suggests that, *by themselves,* ESPPMTs cannot generate quality data for assessing mathematical performance.

So far as reliability is concerned, from our perspective our data indicate that the reported internal consistency index of 0.94 for the *Mathematics Competency Test* suggests that test scores will not only "reliably" generate much helpful information about a student's understandings, *but also much misinformation*—the only trouble is,

without interviewing individual students it is impossible to know what is helpful information and what is misinformation.

Mathematics education research and the use of pencil-and-paper test instruments: Given that recent research studies suggest that about 30% of all responses to questions on ESPPMTs are "mismatches" (Ellerton & Clements, 1995), it appears to be the case that mathematics education research which operationalises and measures a dependent or independent variable through a single application of a pencil-and-paper research instrument is on shaky ground. Since research studies conducted in many countries have relied heavily on such an approach, a review of what constitutes "good design" for quantitative studies in mathematics education research is urgently needed (see Clements and Ellerton (1996) for further discussion of this issue).

In Conclusion

We agree with commentators who fear that pencil-and-paper tests have driven, and continue to drive, other forms of school assessment (Kilpatrick, 1993; Resnick, 1987). Our own research, part of which is reported in this paper, indicates that even the most expertly constructed ESPPMT can never produce an accurate summary of what children really know, unless interviews are part of the data analysis process. Given time constraints, that is most unlikely. The fact that the percentage of mismatches varied considerably between schools in the study reported in this paper suggests that ESPPMTs should not be used for benchmarking school performance.

Acknowledgment

This research was supported by a Large Research Grant (number A79602385) from the Australian Reseach Council.

References

Australian Education Union (1995, Winter). Common cause. *Australian Educator, 7*, 4–6.

Black, P. J. (1994). Performance assessment and accountability: The experience in England and Wales. *Educational and Policy Analysis, 16*(2), 191–203.

Clarke, D. J. (1988). *Assessment alternatives in mathematics.* Canberra: Curriculum Development Centre.

Clarke, D. J. (1996). Assessment. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 327–370). Dordrecht: Kluwer Academic Publishers.

Clements, M. A., & Ellerton, N. F. (1996). *Mathematics education research: Past, present and future*. Bangkok: UNESCO.

Cuttance, P. (1995). Benchmarking in the NSW school system. Unicorn, *21*(2), 60–69.

de Lange, J. (1996). Using and applying mathematics in education. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 49–97). Dordrecht: Kluwer Academic Publishers.

Ellerton, N. F., & Clarkson, P. C. (1992). Language factors in mathematics education. In B. Atweh & J. Watson (Eds.), *Research in mathematics education in Australasia 1988–1991* (pp. 153–178).Brisbane: Mathematics Education Research Group of Australasia.

Ellerton, N. F., & Clements, M. A. (1992). Implications of Newman research for the issue of "What is basic in school mathematics?" In B. Southwell, R. Perry, & K. Owens (Eds.), *Proceedings of the Fifteenth Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 276–284). Sydney: Mathematics Education Research Group of Australasia.

Ellerton, N. F., & Clements, M. A. (1995). Challenging the effectiveness of pencil-and-paper tests in mathematics. In L. Velardi & J. Wakefield (Eds.), Celebrating mathematics learning (pp. 268–276). Melbourne: Mathematical Association of Victoria.

Frary, R. B. (1985). Multiple-choice versus free-response: A simulation study. *Journal of Education Measurement, 22*, 21–31.

Garet, M. S., & Mills, V. L. (1995). Changes in teaching practices: The effects of the Curriculum and Evaluation Standards. *The Mathematics Teacher, 88*(5), 380–389.

Gays, S., & Thomas, M. (1993). Just because they got it right, does it mean they know it? In N. Webb & A. Coxford (Eds.), *Assessment in the mathematics classroom* (pp. 130–134). Reston, VA: NCTM.

Hembree, R, (1987). Effects of noncontent variables on mathematics test performance. *Journal for Research in Mathematics Education, 18*, 197–214.

Holt, M. (1993). The high school curriculum in the United States and the United Kingdom: Perspectives on reform and control. *Journal of Curriculum and Supervision, 8*(2), 157–173.

Izard, J. F., & Miller, K. M. (1996). *Mathematics competency test: User's manual.* Camberwell, Victoria: Australian Council for Educational Research.

Kilpatrick, J. (1993). Beyond face value: Assessing research in mathematics education. in G. Nissen & M. Blomhoj (Eds.), *Criteria for Scientific Quality and Relevance in the Didactics of Mathematics* (pp. 15–34). Roskilde (Denmark): Danish Research Council for the Humanities.

Kindsvatter, R., Wilen, W., & Ishler, M. (1988). *Dynamics of effective teaching.* New York: Longman.

Leder, G. (Ed.). (1992). *Assessment and learning of mathematics.* Hawthorn, Victoria: Australian Council for Educational Research.

Mathematical Sciences Education Board (1993). *Measuring counts: A policy brief.* Washington, DC: National Academy Press.

McGaw, B. (1995). Benchmarking for accountability or improvement. *Unicorn, 21*(2), 7–12.

National Council of Teachers of Mathematics Commission on Standards for School Mathematics) (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics (1995). *Assessment standards for school mathematics.* Reston, VA: Author.

Newman, M. A. (1977). An analysis of sixth-grade pupils' errors on written mathematical tasks. In M. A. Clements & J. Foyster (Eds.), *Research in mathematics education in Australia, 1977* (Vol. 2, pp. 269–287). Melbourne: Swinburne College Press.

Niss, M. (Ed.). (1993). *Investigations into assessment in mathematics education: An ICMI study.* Dordrecht: Kluwer.

Olssen, K. (1993). *Assessment and reporting practices in mathematics.* Canberra: Department of Employment, Education and Training.

Resnick, L. B. (1987). *Education and learning to think.* Washington, DC: National Academy Press.

Ridgway, J., & Passey, D. (1993). An international view of mathematics assessment—Through a class, darkly. In M. Niss (Ed.), *Investigations into assessment in mathematics education: An ICMI study* (pp. 57–72). Dordrecht: Kluwer.

Romberg, T. A. (1993). Assessing mathematics competence and achievement. In H. Berlak, F. M. Newmann, E. Adams, D. A. Archbald, T. Burgess, J. Raven, & T. A. Romberg (Eds.), *Toward a new science of educational testing and assessment* (pp. 23–52). Albany, NY: State University of New York.

Schmidt, M. E., & Brosnan, P. A. (1995). Mathematics assessment: Practices and reporting methods. *School Science and Mathematics, 96*(1), 17–20.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22–27.

Stake, R. E. (1995). The invalidity of standardised testing for measuring mathematics achievement. In T. Romberg (Ed.), *Reform in school mathematics—And authentic assessment* (pp. 173–235). New York: State University of New York.

Thongtawat, N. (1992). *Comparing the effectiveness of multiple-choice and short-answer paper-and-pencil tests.* Penang (Malaysia): SEAMEO/RECSAM.

Vernon, P. E., Miller, K. M., & Izard, J. F. (1996). *Mathematics competency test.* Camberwell, Victoria: Australian Council for Educational Research.

Webb, N. L., & Coxford, A. F. (1993). *Assessment in the mathematics classroom.* Reston, VA: National Council of Teachers of Mathematics.