Chapter 13
Developing Assessment for Learning in a Large-Scale Programme

Hak Ping Tam and Yu-Jen Lu

13.1 Introduction

In Taiwan, as in many school systems around the world, most students are assessed on a regular basis, ranging from low-stakes daily classroom quizzes to high-stakes graduation examinations. The purposes of such assessments vary, with some emphasizing formative and others summative outcomes. As a broad generalization, small-scale assessments lend themselves more easily to assessment for learning than large-scale assessments, which are often designed with the intention of acquiring information about performances in ways that facilitate comparisons across large cohorts of students. As Popham (1999) has pointed out, many large-scale educational assessment programs place too much emphasis on the accountability aspect and too little attention on the instructional aspect. As a result, there is an imbalance between the two functions of assessment as practiced in most programs. Popham suggested that large-scale assessment should instead lend itself more towards improving the instructional practices in the classroom, or at least striking a balance between the two aspects. Yet even if one would decide to heed this suggestion, it begs the question as to how one should take on this challenge and actually carry it out in a large-scale environment. Particularly problematic is the issue of extracting separate yet useful feedback to students, teachers and other stakeholders after analyzing huge amounts of test data from the students involved. There is a risk that such an undertaking might prove to be prohibitively expensive, or that the information thus made available might be too superficial to be of any practical use for instruction.

The purpose of this chapter is to provide a brief introduction to just such a large-scale assessment programme in Yilan County of Taiwan – one that has the dual aim of being supportive towards classroom instruction as well as providing policymakers with the requisite information about students' learning status. This assessment

H.P. Tam (✉)
Graduate Institute of Science Education, National Taiwan Normal University, Taipei, Taiwan
e-mail: t45003@ntnu.edu.tw

R. Berry, B. Adamson (eds.), *Assessment Reform in Education*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 14,
DOI 10.1007/978-94-007-0729-0_13,

programme can serve as an example of what can be done in terms of assessment reform at a county or even higher level. Based on the Yilan experience, this chapter will suggest, in the subsequent discussion, ways to promote assessment for learning within large-scale testing programs in general. Then we will discuss several issues related to the implementation of this kind of large-scale assessment programme and identify some possible directions for future development.

Yilan is a county on the northeastern coast of Taiwan with a sparsely distributed population of about 460,000 people. Since 2006, the Yilan County has been carrying out a county-wide assessment programme, in accordance with the requirement from the Ministry of Education that local governments should monitor the progress of their students by means of assessment (Ministry of Education, 2003). This chapter will describe the practices in the year 2007 as an example. A guidance committee with two staff was commissioned to oversee the planning of the mathematics assessment programme for the fourth grade. This committee recognized that most assessments performed in schools on a routine basis are achievement oriented. In order to be innovative, they established the ambitious goal at the outset that the assessment for fourth grade mathematics should be designed with the intention of providing useful information to the mathematics teachers, school principals as well as the parents of the participating students.

13.2 The Assessment Framework

In Taiwan, the formal version of the official Grade 1-9 Curriculum Guidelines was introduced in 2003 and was implemented through several stages to cover all grades in primary and junior high schools. Accordingly, the guidance committee decided that the content of the mathematics assessment should be fully aligned with the standards specified in the official curriculum. A committee of 15 local expert mathematics teachers was assembled to compile the assessment framework for the whole county. Each member of the committee had been recognized as an expert teacher in mathematics at the primary school level. Their average experience of teaching mathematics amounted to about 12.2 years.

Since there is a wide array of standards specified in the official curriculum, the first task of the committee was to decide on the content and the cognitive domains for the assessment framework. Their work could be summarized as a three-step procedure. Firstly, the committee studied and analyzed the official mathematics curriculum standards in detail. Additional reference was made to both the Trends in International Mathematics and Science Study 2003 and the US National Assessment of Educational Progress 2005 assessment frameworks. Secondly, the committee had to identify important standards in the curriculum to be tested. Since the county's education officers had decided to allot just one regular class period for the assessment, the committee recognized that only a few curriculum standards could be tested within such a short duration of time, given that there are thirty two curriculum standards that are stipulated for the fourth grade students to learn. In order to identify the appropriate set of standards that would form the basis of the test, the members

were asked individually to rate the importance of each standard on a 5-point Likert scale. Those standards with the highest total ratings were prioritized for inclusion in the test. The final criterion for selection lay in the essentiality, interpretability, assessability and the richness of mathematical content associated with the standards as judged by the majority of the committee members. Thirdly, they took practicality into consideration to narrow down on their choices. In view of a limited budget and the vast number of students who would participate in the main survey, they decided that only multiple choices test items would be used to assess the students. The committee further decided to restrict the whole instrument to 30 multiple-choice items so as to fit into the limited accessibility of testing time. Under this constraint, they further trimmed down the scope of the assessment to encompass ten curriculum standards that spanned over 16 fundamental topics.

The committee further resolved that the content domain should include the following five areas: number and computation, quantity and measurement, geometry, algebra, as well as probability and data analysis. For the cognitive domain, conceptual comprehension, procedural knowledge and problem solving were chosen to be the three essential components. The finalised assessment framework is shown in Table 13.1 below.

Table 13.1 The assessment framework for the Yilan County's fourth grade mathematics test

Content domain	
Number and computation	40%
Quantity and measurement	33%
Geometry	13%
Algebra	7%
Probability and data analysis	7%
Cognitive domain	
Conceptual understanding	33%
Procedural knowledge	46%
Problem solving	21%

13.3 Dual Items Design

In order to facilitate a formative aspect in the county-wide assessment, the committee incorporated a special design in its construction of the test instrument. Since the purpose was to obtain rich information of student performances, the committee wanted to avoid constructing a test with a mere compilation of total test scores. Furthermore, the multiple choice item format has a well known weakness in that students can by chance guess the correct answer. If Yilan County would like to find out the actual status of their students with respect to the 16 fundamental mathematical topics, something needed to be done to enrich the multiple choice format. Towards this end, it was decided that two parallel items that tested basically the same concept should be employed for 14 of the topic areas, accounting for a total of 28 items. The remaining two items were not paired due to the limitation of testing time. However, they were both related to interpreting statistical graphs. One important characteristic

of this design lies in the fact that the distractors across the dual items correspond to each other. In other words, the distractors were set up such that the same set of misconceptions was being used as options in both items.

The main reason for using this design is to control for chance performance by the students. Obviously, if a certain student displayed the same misconception in the item pair, one could be more confident in identifying the student as holding that misconception. In contrast, if students could consistently get both items correct, there is a high chance that they had already grasped the concept being tested. However, if some students picked the right answer for one item but got the other item wrong, this might indicate that they either did not have a firm understanding of the concept, had misconceptions or had been lucky in guessing a correct answer. One other piece of valuable information would come from those students who displayed different misconceptions across the two items. This probably revealed that they had only a fuzzy understanding of the concept being tested or had been guessing randomly in their responses.

The committee finalized the instrument after four rounds of field testing. The Education Office of the Yilan County decided that in 2007, every fourth grade student in the county should participate in the assessment programme. This amounted to a total of 6,374 fourth grade students from 76 primary schools participating in the mathematics assessment. In order not to impede the regular course of instruction, the assessment was administered in late June of 2007. The test results were analyzed via descriptive statistics, classical test theory and the Rasch modeling approach. They are not reported here due to the limitation of space, and interested readers can refer to Tam and Lu (2008) for more details. It is only mentioned in passing that the test data substantiated very decent reliability coefficients, with KR-20 at 0.88 and split-half at 0.95. Misconceptions about various mathematical topics could be identified. The information was disseminated to the teachers by means of a specially designed report system.

13.4 Online Report System

A tremendous amount of effort was invested to make the test results from the Yilan's assessment programme available to all concerned parties via the internet. The intended list of information recipients included the county's officials, school principals, school mathematics teachers, participating students and their parents as well as the general public. An elaborate online report system was set up so that different amounts of information were made accessible to various parties according to their level of authorization. The purpose behind such an investment is to optimize the applicability of the analysis results so as to facilitate subsequent instructional and/or remedial endeavor (Tam & Lu, 2008). Figure 13.1 displays the front page of the report online system. The school site is password protected and reserved for access by teachers and school officials. The public site is open to any interested parties. Parents can, in addition, use passwords to check the performances of their children. The public site also includes links to webpages on the results of other school subjects being tested. All the webpages are written in Chinese. In this chapter,

Fig. 13.1 The front page of the online report system

the webpages that follow were taken from the report by the Education Department of Yilan County (2008) and translated into English for international readers.

This system allows designated school officials and every fourth grade mathematics teacher to examine the performance of their students at the individual level as well as at the class and school levels. The results of students' performance are organized in various ways for different purposes. For example, there are webpages on students' performance according to the content and cognitive domains as specified in the assessment framework (see Fig. 13.2).

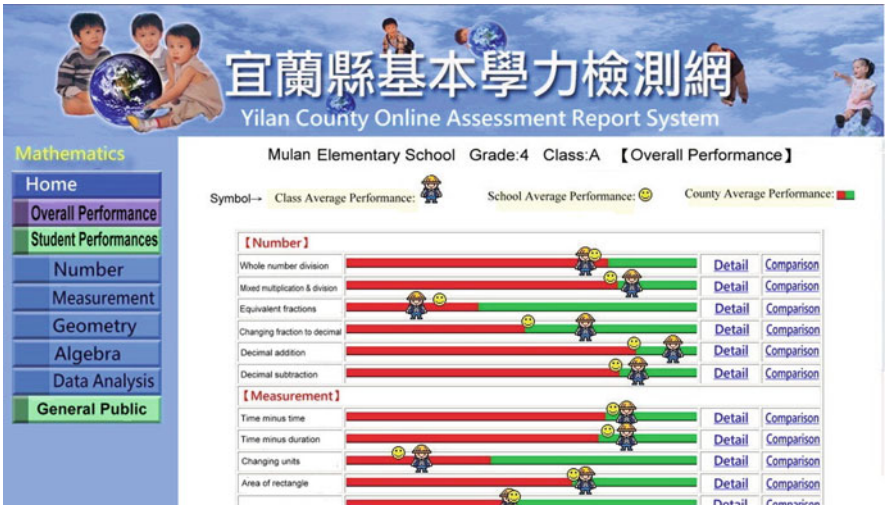

Fig. 13.2 A sample webpage illustrating the overall mean performances for a particular class of students in relation to the school and county averages (NB. The data shown here is for illustrative purpose and is not real data.)

Figure 13.2 shows a sample webpage from the online system for teachers displaying the mean performances for a particular class of students on various topics as compared to the average school and the overall county performances. More detailed results are also presented according to each curriculum standard being tested. Fourth grade teachers can find useful information regarding the types of misconceptions their students might have. There are further links to detailed reports as well as comparisons with the performances of other schools of similar size. Figure 13.3 shows a sample webpage that delineates the distribution of misconceptions across an item pair on a specific topic for a particular class of students. The numbers within each cell indicate the frequency count as well as the percentage of students with the corresponding combination of responses across the item pair. Moving the mouse over any cell would invoke a small pop-up window that reported the seat number of those students who were classified into that cell. In addition, the teachers could click on the hyperlinks of the three misconceptions listed in the headings to read more about their features together with a few suggestions about how to handle them in class.

In addition, school officials could examine their school's performance in comparison to the average performance of all schools across the whole county or the town in which the school is located (see Fig. 13.4). There are webpages that display the average performances of individual school in various subjects. Furthermore, standards setting procedures are used to classify students according to their levels of proficiency in various subjects. In fourth grade mathematics, three categories are used: "effort needed", "met the standard" and "excellent".

Item2 / Item1	Correct	Misconcept A	Misconcept B	Misconcept C	Others
Correct	14 46.67%				
Misconcept A			1 3.33%	3 10.00%	
Misconcept B			9 30.00%		
Misconcept C		2 6.67%			
Others				1 3.33%	

ID: 2,8,11,12,16,20,25,26,31

Fig. 13.3 A sample webpage illustrating the distribution of performances across a dual pair of items with the IDs of students who committed errors typical of misconception B shown

Region	School			City			County		
Value	Effort needed	Met Standard	Excellent	Effort needed	Met Standard	Excellent	Effort needed	Met Standard	Excellent
3rd CHN	16.88%	51.30%	31.82%	26.28%	50.76%	22.97%	35.89%	46.16%	17.95%
4th Math	12.50%	54.17%	33.33%	11.74%	49.08%	39.17%	18.78%	50.02%	31.20%
6th CHN	14.14%	42.42%	43.44%	15.08%	40.78%	44.14%	20.62%	42.39%	36.99%
6th ENG	23.87%	24.90%	51.23%	27.94%	21.97%	50.09%	34.63%	25.04%	40.33%
6th Math	26.23%	38.73%	35.04%	24.97%	43.87%	31.16%	30.18%	44.22%	25.60%

Fig. 13.4 A sample webpage from the online report system for school officials that illustrates the average performances of a school on various topics

A special report card was designed for the dissemination of assessment results to the students and their parents. Figure 13.5 shows a portion of the individualized student report card. The information is relatively easy to interpret and avoids direct referral to grades. The main purpose of this report card is for communication rather

Items	Topics	Performance	Comment	Overall
1	Whole number division	☆☆☆	You did real well in dividing 3-digit numbers by 2-digit numbers.	
2	Mixed multiplication and division	☆☆☆	You did real well in computing mixed multiplication and division problems.	★★
3	Equivalent fractions	☆	You may have mistaken 2/5=2/7, 3/4=3/7. (See misconception 2)	
4	Interchange between decimal and fraction	☆	Please pursue further understanding of the interchange between decimal and fraction. (See misconception 2)	Note from school
5	Decimal addition	☆☆☆	You did real well in addition problems involving 3 decimal places.	
6	Decimal subtraction	☆☆☆	You did real well in subtraction problems involving 2 decimal places.	
7	Time point minus time point	☆☆☆	You did real well in problems involving time point minus time point.	
8	Time point minus duration	☆☆☆	You did real well in problems involving time point minus duration.	
9	Changing units of area	☆	You may need clarification in square centimeter and square meter.	
10	Perimeter of rectangle	☆☆☆	You did real well in items involving perimeter of rectangles.	Comment from parent
11	Area of rectangle	☆	You may have misapplied the area formula. (See misconception 2)	
12	Perpendicularity	☆☆☆	You understand perpendicularity real well.	
13	Parallelism	☆	Please pursue further understanding of what parallel means.	
14	Priority of operations	☆	You may have only one strategy to solve items on priority of operations.	
15	Statistical graph	☆☆☆	You did real well in reading off information from stat. graphs.	
Signatures		Principal	Dean of Studies	Teacher	Parent

Fig. 13.5 A sample report card of students' performance with information for students and parents

than for comparison. It is designed to help students to realize which areas they need to strengthen. Nevertheless, parents could have an idea about their children's performance through a 3 white-star rating scale that is used for each of the 15 topics being assessed. The report card also contains a 3 black-star scaling system that represents the overall performance of the student on the whole test. In both scales, top performance is indicated by three stars while poor performance receives only one star. The adoption of the starring system is to promote the diagnostic aspect of the test and attenuate the tendency of parents to compare their children's performances with others. Finally, there are concise verbal descriptions of the students' performances on each topic. Praise for good performance as well as diagnostic suggestions are expressed in a conversational manner.

Given the novelty of the ideas behind the assessment programme together with the intention of maximizing the usability of the assessment results, two workshops were organized for local teachers the attendance of which was mandatory. Instructional procedures were presented during the workshops so as to familiarize the teachers with navigating through the website. The various features of the report system were also explained. After comprehensive demonstrations, a time was set aside for discussions and questions. Important misconceptions were clarified and discussion was initiated with respect to how teachers could best use the test results to plan their remedial activities for these misconceptions as well as other teaching practices. Since the report system is quite elaborate and quite unlike anything instituted in the county before, actual demonstrations of the operations of the system had to be organized for all the fourth grade teachers. Such training sessions were deemed an essential component of the assessment programme by the committee.

13.5 Issues to Consider

The Yilan County's assessment programme represents an attempt at improving the support for classroom instruction by way of providing teachers with diagnostic information about the learning status of their students on important mathematical topics. Another purpose of the assessment is to inform the parents about the learning status of their children. Special effort was invested in compiling an informative report card that aims at providing direct attention to potential misconceptions that a student may have (Sadler, 1989), thereby facilitating remedial education at home. The whole operation demonstrates that assessment for learning is feasible even at the level of a large-scale assessment programme. However, a number of issues and challenges arose in the implementation of the programme.

The adoption of the dual items design is a special feature that helps to render the county's assessment into an assessment for instruction. This idea was not easily accepted at first, since it differs from the experiences encountered by most of the local teachers. It took some time for the idea to precipitate. In Taiwan, most assessments take the form of an achievement test, in which a wide spectrum of topics as listed in the curriculum should be represented. Accordingly, each item can

only focus on testing one topic. Thus it appears on the surface that the dual items design represents a waste of resources. Another apparent limitation of the design is the small number of topics (only 16 out of 40 topics) being assessed. Moreover, the timing of the assessment is also an important matter of concern. The choice of June meant that it was conducted close to the end of the school year in Taiwan, given the requirement that it should not interfere with the regular schedule of instruction in schools. As a result, the analysis as well as the compilation of results was carried out during the summer break and the dissemination of results took place at the beginning of the subsequent academic year. Thus contrary to the purpose of formative assessment, borrowed from Scriven's (1967) concept of formative evaluation, the results from the appraisal could not be put to immediate use by the teachers. In this sense, the timing of the operation made it appear more like a summative assessment than a formative one. Two concerns arise here. First, after returning from the summer break, the students may have forgotten part of what they learnt in grade four. Second, in most cases the students will not have the same teachers teaching them mathematics in grade five. It will be less effective if a new teacher has to take care of students' misconceptions in mathematics if he or she has not taught them before. However, this disadvantage has more to do with the administrative constraints that were set by the county office rather than with an inherent flaw in the design.

Another obstacle to successful implementation has to do with reservation, or even skepticism, from some of the teachers because the assessment was initiated by the Education Department of Yilan County and was a top-down operation. Without a sense of ownership, some teachers were suspicious about its real purposes and felt threatened that they would be evaluated by how well their students performed in the county-wide assessment. The low stakes status of the assessment did not help much in desensitizing the apprehension of these teachers. Some of the information presented in the online report system might have actually led the teachers to suspect that they were also being assessed within the operation system. For example, in the online report system, there was information in a graphical display that compared the average performance of each class with respect to the average performances of all the fourth grade classes within the same school and also with respect to the average across the whole county. Some teachers might have interpreted this information as a way of comparing their effectiveness with other mathematics teachers within the same school or within the same county. Likewise, school principals might reason along the same lines and become anxious about the performance of their schools with respect to the other schools in the county.

A further undesirable outcome is that it may initiate a tendency for some teachers to teach to the test. At first, this may seem unlikely as the assessment is a low stakes test in Yilan. Yet from another angle, the likelihood is actually not negligible because all the items were made accessible to the teachers after the assessment. Moreover, the misconceptions associated with each option were also explained and displayed in the online report system. As a result, there is ample opportunity for some conscientious teachers to prepare similar items and then teach them to their students rather than focusing on the mathematical concepts. Nevertheless, it is still too early to tell whether this will happen in Yilan, since the current assessment programme and the

report system were set up only quite recently. Follow-up observations are essential in this regard.

The decision to disseminate the results through a report card that is specially designed for the parents and the students could also be regarded as controversial. On the one hand, the feedback was seen as valuable for the parents and students, who can take advantage of the information provided and try to improve in their knowledge. On the other hand, it could be argued that the results might have an adverse impact. For example, the results might become a source of unnecessary competition among the students (Black & Wiliam, 1998a, 1998b), and the report card might turn into a source of over-concern for the parents towards the performance of their children. Although the report card included diagnostic statements with explicit suggestions so as to attenuate the tendency by parents to compare their children's performance with others', the practice of using stars to indicate the level of student performance in the assessment programme can still be regarded as a kind of grading system.

13.6 Future Directions for Improvement

This chapter will close by suggesting some possible future directions for enhancing the assessment programme of Yilan County. First and foremost, it is suggested that the reservations and concerns of various stakeholders towards the programme should be taken seriously and addressed. The active participation of teachers in the project should be encouraged. One way of increasing the level of teacher's involvement in the data dissemination phase of the assessment is to open up various channels through which expert teachers can share effective ways of handling the listed misconceptions by fourth grade students. Also, mutual communication should be established and measures should be undertaken to alleviate their worries that the programme serves as an implicit way to perform teacher evaluation. More consideration needs to be devoted towards devising a meaningful grading system as well as an informative report system, so that the intentions of the programme can be more easily appreciated by teachers, school principals, students, parents and education officials. A useful reference on grading and reporting is discussed in the article by Brookhart (1999), even though it is basically written for pre-service teachers. The section on setting meaning for grades can also have implication for large-scale assessments. Meanwhile, a more extensive study should be conducted to evaluate the success of the whole assessment programme, especially in relation to how well the stakeholders perceive the efficiency and the effectiveness of the programme. The educational impact that has been generated should also be gauged whenever possible.

In terms of the technical design of the assessment, some adjustments might be required. The suitability of using Rasch modeling to analyze the data should be reconsidered. Because of the special dual item structure adopted in the formal instrument, the relationship between the item pair would very likely violate the local

independence assumption required by item response theory. It is suggested that more advanced techniques, such as the testlet response theory (Wainer, Bradlow, & Wang, 2007), that can handle this violation should be adopted to re-analyze the 2007 data as well as the data in the years to come. Another technical matter concerns the establishment of formal standard setting procedures for the assessment programme. Currently, there is a three black-star system on the report cards that reflects the overall performances of the participating students. However, this system was set up by consensus among members of the administrative team rather than by using more rigorous methods. It is suggested that the current system should be removed from the report cards in the future. In other words, only the performance level with respect to each mathematical topic should be provided to each student. This would be more in tune with the purpose of assessment for learning and instruction. On the other hand, should the overall performance decision be deemed essential, formal benchmarks corresponding to the basic, proficient and advanced level of performances should be established using appropriate standard setting procedures. Meanwhile, education officers in the county can set up goals for their students' attainment in mathematics with respect to these standards. For example, they can establish the goal that 80% of all students in the county should reach the proficiency level. These standards can then provide incentives and directions for improvement should the students' performances fall short of this goal. Also, these goals can illuminate the need for adjustment on the existing educational policies or practices at the school or even at the county level. In sum, the intention of these standards would be to bring about academic success to as many students as possible.

Administratively and pedagogically speaking, the timing of assessment is an important consideration. It would probably be better to carry out the assessment at a date well before the end of the school year so that the teachers have ample opportunity to modify their teaching plans and incorporate the diagnostic information into their instructional practices. Since in many cases, the fourth grade mathematics teachers may be different from the fifth grade teachers, an earlier date can avoid the embarrassing situation of requesting the upper grade teachers to deal with the misconceptions of students accrued under the tutorship of other teachers (Black & Wiliam, 1998b).

While the innovative use of assessment for learning in the large-scale programme outlined in this chapter has shown promising signs of success, it currently is only administered in fourth grade mathematics. Implementing it in other subject areas and at other grade levels in the future might create new sets of challenges. Such expansion, while seemingly desirable, does pose another danger – that the programme will become a victim of its own success. More large-scale assessments might be viewed as desirable, or even the norm, thereby resulting in more intrusion on instruction and demands on more time for testing. As a result, the scope for using small-scale formative assessment in regular classrooms – which, it could be argued, is more desirable than anything large-scale assessment can offer (Sadler, 1989) – might be diminished.

Howard Wainer (2007) argues that, while there has already been much progress in the field of psychometrics, further developments should concentrate on getting

"the bang out of the buck" from tests rather than on test theory. The assessment programme that took place in Yilan represents an attempt along this line of thinking by endeavouring to make the assessment more useful to both teachers and students. Of course, there is no quick fix in educational matters and many challenges still remain. However, the attempt to transcend the traditional role of large-scale assessment towards that of assessment for learning and instruction represents one small step of assessment reform in Yilan.

Disclaimer The opinions expressed in this paper are those of the authors and do not necessarily represent the position of the Education Department of the Yilan County.

References

Black, P. J., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7–74.

Black, P. J., & Wiliam, D. (1998b). Inside the black box: Raising standards through educational assessment. *Phi Delta Kaplan*, *80*, 139–144.

Brookhart, S. M. (1999). Teaching about communicating assessment results and grading. *Educational Measurement: Issues and Practice*, *18*(1), 6–13.

Education Department of Yilan County (2008). *2008 Yilan county assessment: Report in language arts and mathematics*. Yilan, Taiwan: Education Department.

Ministry of Education (2003). *General grade 1–9 curriculum guidelines for elementary and junior high school education – Area of mathematics learning*. Taipei, Taiwan: Author (in Chinese).

Popham, W. J. (1999). Where large scale educational assessment is heading and why it shouldn't. *Educational Measurement: Issues and Practice*, *18*(3), 13–17.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–144.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. W. Gange, & M. Scriven (Eds.), *Perspectives of curriculum evaluation*. Chicago: Rand McNally.

Tam, H. P., & Lu, Y. J. (2008). *Assessment of students' performance in mathematics at the county level – The Yilan County's approach*. Paper presented at the eleventh International Council of Mathematical Education in Monterrey, Mexico.

Wainer, H. (2007). A psychometric cicada: Educational measurement returns. *Educational Researcher*, *36*(8), 485–486.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.