Chapter 12
Using User-Defined Fit Statistic to Analyze Two-Tier Items in Mathematics

Hak Ping Tam, Margaret Wu, Doris Ching Heung Lau, and Magdalena Mo Ching Mok

12.1 Introduction

The two-tier item is a relatively new diagnostic item format for classroom assessment and is gradually gaining popularity in certain areas of educational research. For the past two decades, it has been used to assess at a deeper level students' understanding of the concepts being covered in classes, especially in the area of science education (e.g., Treagust 1988; Treagust and Smith 1989; Tan and Treagust 1999). Its popularity in recent years may be partly illustrated with the following piece of information. In 2007, a whole issue of the *International Journal of Science Education* was devoted to reporting the research design and results of a study entitled National Science Concept Learning Study (NSCLS). This study was conducted in Taiwan in 2003 and involved more than 30,000 students from primary to senior high school

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

M. Wu
Work-based Education Research Centre, Victoria University, Melbourne, Australia
e-mail: wu@edmeasurement.com.au

D.C.H. Lau
Formerly Centre for Assessment Research and Development,
The Hong Kong Institute of Education,Tai Po, Hong Kong

The University of Hong Kong, Hong Kong
e-mail: chlaudoris@gmail.com

M. Mo Ching Mok
Department of Psychological Studies, and Assessment Research Centre,
The Hong Kong Institute of Education, 10 Lo PingRoad, Tai Po, N.T., Hong Kong
e-mail: mmcmok@ied.edu.hk

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_12, © Springer Science+Business Media Dordrecht 2013

(Guo 2007; Tam and Li 2007). Its main purpose was to assess students' misconceptions of important science concepts from primary to senior high school. What is worth noticing is that all the items adopted in this study were framed in a two-tier format.

A two-tier item can be viewed as a special kind of testlet in that it has a common item stem followed by two subitems, with one of them requiring the respondents to carry out part of the task while the subsequent subitem requiring them to finish the remaining part of the task. In science education, a typical two-tier item is made up of an item stem followed by two portions. Usually, the purpose of the first portion is to assess whether students could identify some factual aspects with respect to a phenomenon stated in the item stem, while the second portion examines if they can supply the correct reason associated with why the phenomenon occurs. Since some students may not be able to identify the correct option associated with the first portion, what they chose as the accompanying reason in the second portion could then reveal valuable information about their knowledge status about the phenomenon being tested. More specifically, the combination of options being chosen across the two tiers has the potential of revealing the misconception being held by students about why some phenomenon happens or does not happen. As a result, this item format has been used as one of the ways to illuminate the kind of misconceptions as well as how widespread they have been among the students taking the test.

Although this format has not quite found its way into research conducted in the area of mathematics education, there are, nevertheless, situations where some mathematics items can be essentially treated as two-tier items. For example, one common way of assessing students' abilities in solving word problems is to ask them to formulate an equation that corresponds to the conditions given in the items. In some tests, partial credits will already be assigned to examinees who have been successful in expressing the correct equation. Afterwards, the examinees are required to solve the equations they have formulated and then provide their final answers. Again in some tests, partial credits may be assigned to those who can provide the correct final answer. Thus accordingly, one can view a word problem as the item stem and the requirement to set up the corresponding equation as the first tier while the compilation of the final answer as the second tier. As a matter of actual practices, this approach has been frequently adopted by mathematics teachers especially in the elementary grades.

Unfortunately, the methodology regarding how the two-tier items should be analyzed is still fairly underdeveloped in the area of science education. For example, many data analysis, as can be currently identified in the literature, is limited to reporting tables of percentages of options being chosen by the examinees across the two tiers for each individual item. This approach is descriptive in nature and is dependent on the sample of students taking the tests. The quality of a two-tier item, moreover, is usually assessed by appealing to the judgment of subject matter experts based on their professional experiences. However, there are many a time when professional judgment cannot be easily made, such as when the two-tier item format appears brand new to the experts. In real practices, it is quite often the other way around with the subject matter experts requesting the data analysts to provide them with supportive statistical information, thereby assisting them in their judgment making regarding whether the two-tier items are in good shape.

One possibility is to use the techniques that have been developed for analyzing testlets or item bundles as reported in the literature. One such alternative is the testlet response theory developed by Wainer et al. (2007). This theory is accompanied by a software program entitled Scoright, which is freely available by way of Educational Testing Service (ETS), thereby making it more attractive to applied researchers. Yet, the technicality behind the testlet response theory is quite involved for most school teachers or even applied researchers to comprehend. In addition, the current version of Scoright is not as user friendly as one would desire. Furthermore, since the theory is based on the Bayesian approach in its estimation of parameters, Scoright can be quite slow in terms of program execution. Though the program allows starting values to be provided by the users so as to speed up the estimation process, many school teachers or applied researchers may find it difficult and need help in deciding on a good set of starting values. There can also be times when the program cannot converge at all in its execution. Thus, it seems that a friendlier approach is much desirable for the common practitioners so that they can handle the analysis of two-tier items in an easy-to-understand manner. Since such information is currently unavailable at large, thus there appears to be a need in developing useful technique for analysis that takes into consideration the relationship between the two tiers within the same item.

12.2 Purpose of Study

A three-step procedure has been proposed in Tam and Wu (2009) as an all-purpose approach to analyze two-tier items. Such practical information as the scoring of the item, the dependence between the two tiers, as well as the functioning of the items can then be provided to the item writers for item evaluation and revision. Since the third step is similar to the item analysis procedure that is commonly seen in a Rasch analysis setting, this chapter aims at illustrating the first two steps that are particularly important for the two-tier item format. More specifically, this chapter will first occupy itself with assessing if there are dependencies between the two tiers for each item on the test as one would expect from the nature of this particular format of item. Afterwards, this chapter will turn its attention to investigate how two-tier items should be scored in the first place. These two steps of data analysis are especially relevant to the data set from a mathematics test with a two-tier structure which will be used to demonstrate the procedure discussed herein. Both the method with its rationale and the data employed for demonstration will be described in more detail in the next two sections. They will then be followed by the results section. Finally, the specific issue about whether all two-tier items should be scored the same way together with other issues of more general interest will be dealt with in the discussion section.

12.3 Method

As a start, one useful yet succinct way of organizing the overall performances of students with respect to a typical two-tier item is to construct a two-by-two cross-tabulation table for the distribution of the students' proportions of right or wrong across the two tiers as illustrated in Table 12.1 below. Among the students who sit for the test, let x be the proportion of those who got both the factual and the reason portions correctly. Similarly, let y be the proportion that got both portions wrongly, z be the proportion that got the factual portion correctly but the reason portion wrongly, and w be the proportion that got the factual portion wrongly but the reason portion correctly.

The original data analysis procedure proposed by Tam and Wu (2009) was comprised of three steps, each tapping into a different kind of information from the two-tier items that appear on the test. The rationale behind this procedure is as follows. Since both tiers, by nature of the item structure, access the same piece of information in the item stem, it is regarded as being safe to assume that students' performances with respect to the two tiers will be related to each other. Hence, the purpose of their first step is to discern systematically if there exists a dependency between the two portions for each two-tier item. If it so happened that the dependency between the two tiers is found to be low for some items, reasonable doubt could then be raised concerning whether these items have functioned according to the intent of the item writers. These items should either be deleted or subjected to revision by referring them back to the item writers. After the relationships between the two tiers have been established, one can then consider how the items should be scored. For example, should the data analyst score the items by using partial credits or should the item be considered correct only when both portions are answered correctly? If the items were inappropriately scored, then any subsequent effort in item analysis and interpretation of results would most likely be led astray. Hence, the second step of the proposed procedure will concentrate on selecting an appropriate item response model that can take into account the dependencies between the two tiers. It is deemed essential to notice that even for those items with justifiable relationship across the two tiers, they have to be properly scored with an appropriate item response model before further item analysis on the items be performed. The main concern of the third step is the provision of item information to subject matter experts that may be useful for revising and rewriting the items. These steps are explained in more detail as follows.

Table 12.1 Proportions of students' performance across the two portions of a two-tier item

Second tier First tier	Right	Wrong	Row total
Right	x	z	$x+z$
Wrong	w	y	$w+y$
Column total	$x+w$	$z+y$	1

Tam and Wu (2009) pointed out that if dependencies exist across the two portions within the two-tier structure, the local independence assumption behind the item response modeling approach will in principle be violated should an item response model be attempted on the data (Embretson and Reise 2000). In order to detect this violation, the user-defined fit statistic as discussed in Adams and Wu (2011) can be applied. The gist of this test statistic will be explained here briefly. Let us first consider an examination that is made up of the usual multiple-choice test items. The user-defined fit statistic allows the data analyst to define several items as a group. The number of items correct within the group is then counted for each participant, which can be regarded as the sum score obtained by each participant. If there is no violation of the local independence assumption within the group of multiple-choice items in the first place, then the sum score should also fit the item response model. However, if there is dependency among the items in a group, then the sum score will tend to be too high or too low than expected, owing to the relationship among the items in the group. Thus, when all the items satisfy the local independence assumption except for those items defined in the group, this group of items can be picked up by the user-defined fit statistic as not fitting the item response model applied. The user-defined fit statistic is implemented in the ConQuest and can be used to compute any groupings of items in a test (Wu et al. 1998). The sum score can be tested against the sum score of yet another group of items also defined by the data analyst. This idea can then be extended to the situation when the multiple-choice test has an extra two-tier item added. For this particular two-tier item, if a respondent scores high on one tier, then it is likely that the same respondent will also score high on the other tier. Thus, when the two portions of the two-tier item are treated as a group, it can be picked up by the user-defined fit statistic. In this chapter, a data set will be used to demonstrate the procedure discussed herewith. This data came from an examination that consisted of both regular items and a number of two-tier items. In our first analysis, an item response model was fit to the two-tier items as if they were all made up of independent items. In other words, the relationships between the two tiers were ignored in this round of analysis. Fit statistics were then computed for the two tiers in each item pair. The magnitude of the fit statistic provides a measure of model violation, thereby revealing how closely the two tiers are related within each item pair.

After the relationship between the two tiers has been established, the second step focuses on selecting an appropriate item response model that can account for the dependencies between the two tiers of the items. For example, each two-tier item can be modeled as one item by scoring it by ways of assigning partial credits. In this step, the data analyst should consider a number of item response models that might reasonably be used to score the two-tier items. These models will then be applied to the data. Model comparisons are made by means of the model fit statistics as well as the test reliability information from each model. The best fitting model could then be used for calibrating and further purposes.

The third step involves the extraction of information at the level of response categories so as to assist item writers in assessing how each item pair functioned. In addition to the frequencies or proportions of respondents in the various response categories, the average ability is, for example, also useful information, as well as the corresponding item characteristic curves by category. As explained earlier, the third step is more familiar to applied researchers, so this chapter will focus on delineating the first two steps. The data set that will be utilized to illustrate the suggested procedure will be described in the next section.

12.4 Data

The aforementioned methodology has been applied to a set of data collected in 2010 by the Assessment Research Centre of The Hong Kong Institute of Education. The test instrument was made up of ten mathematics items cater for students at the fifth-grade level in Hong Kong. The contents being tested included eight items on fractions, one item on rearranging a given set of digits to obtain the smallest number, and one item on finding the greatest common divisor out of a given set of numbers. Of the eight items related to fractions, two of them were purely computational type of items while the other six were word problems. Furthermore, three of the word problems required the respondents to list out their steps before reporting their answers. These three items were regarded as two-tier items in the present study. The other items only required the respondents to write down their answers. A total of 860 fifth-grade students participated in the test. This data set was actually part of a larger study, the purpose of which did not affect in whatever way the methodology proposed hereby in this chapter. All the analysis was performed by using the specialized software program ConQuest (Wu et al. 1998).

12.5 Results

In order to explore if the three two-tier items really did violate the local independence assumption, the step listing portion and the corresponding answer reporting portion were treated as a group for each item. They were labeled as items 7.1, 7.2, 8.1, 8.2, 9.1, and 9.2, respectively. User-defined fit statistics were applied to these three two-tier items. For comparison purpose, individual portion of the three two-tier items were treated as independent items and were randomly paired with the rest of the items on the test form. User-defined fit statistics were also applied to each of these random pairs of items. The results were reported in Table 12.2 below. The user-defined fit statistic in Table 12.2 can be regarded as approximately a *z*-test. When its value is within the range of −2 and 2, the item pair could be regarded as fitting the item response model that is based on the assumption of local independence. When the user-defined fit statistic is greater than 2, the item pair is regarded as having a dependency beyond what the item response model assumes. When it is less than −2, the item pair is regarded as testing different constructs.

Table 12.2 User-defined fit statistics for the two-tier items that were correctly paired and also for items that were randomly paired

Portions of two-tier items correctly paired		Items paired randomly	
Item pair	Fit statistic (weighted)	Item pair	Fit statistic (weighted)
7.1, 7.2	12.162	1, 7.1	–0.672
8.1, 8.2	13.203	2, 7.2	–0.408
9.1, 9.2	13.693	3, 8.1	–1.592
		4, 8.2	–1.830
		5, 9.1	–0.606
		6, 9.2	–2.158
		1, 2	3.161
		3, 4	1.417
		5, 6	1.909
		8.1, 10	–0.087
		8.2, 10	–0.111

It can be seen that when the portions from the two-tier items were correctly paired, the three two-tier items (i.e., items 7–9) all had a user-defined fit statistic much greater than 2, thereby indicating that dependencies existed between the two portions within each two-tier item. In contrast, for items that were randomly paired, as reported in the last two right-hand columns in Table 12.2, their user-defined fit statistics were mostly within the range from −2 to 2. This indicated that the item pairs did not have a dependency over and beyond what the item response model assumed. This was especially the case for both items 7 and 8. When the portions from these items were paired up randomly with other items in the same test, their user-defined fit statistics tended to be much smaller in magnitude. Another interesting observation is that the first portion of each two-tier item tended to have a smaller, at least in a relative sense, user-defined fit statistic when they were randomly paired with the non-two-tier items. This may be attributable to the fact that the computational step of a mathematics item will under most circumstances be unrelated to the answer of another mathematics item. It must be emphasized that the pairing of one portion from a two-tier item with a non-two-tier item is not limited to those listed in Table 12.2. The pairings listed over there are for demonstration purposes. Should any doubt ever arise, another round of random pairings can be pursued and the user-defined fit statistics performed again. However, one should be careful not to capitalize on chances by running too many tests.

It was further noticed that not only did the three two-tier items demonstrate a similar pattern in terms of high user-defined fit statistics, they also shared a similar distribution of respondents’ performances in proportions with respect to the two-tier structure. As a typical example, Table 12.3 reported the distribution of participants’ proportions across various combinations of right and wrong with respect to the two-tier structure in item 7.

As can be seen from the table, the majority of the respondents (almost 95%) answered either correctly or incorrectly to both tiers of item 7 at the same time. There were a few respondents who had written down the computational portion

Table 12.3 Distribution of respondents' performances in proportions with respect to item 7

First tier \ Second tier	Right (%)	Wrong (%)
Right	66.16	5.47
Wrong	0	28.37

correctly but yet provided the wrong answer. However, there were no respondents who could obtain the correct answer for the second tier yet missed out on the first tier. These findings make empirical sense since the first tier of this item required the respondents to write down the expression that was necessary for computing the answer. Logically speaking, one must first get the computational portion correct before one can obtain the right answer to the item. It is highly uncommon that a wrongly formulated expression would still render the right answer in real life situation. Apparently, the dependence between the two tiers is fairly strong as demonstrated by the distribution of respondents' performances with respect to this item. After taking all the information together, it can be regarded that the two-tier structure of this particular mathematics exam has been substantiated by results from the user-defined fit statistic.

The second step of the suggested procedure in Tam and Wu (2009) involved the selection of an appropriate item response model. Four separate models were fitted to the group of two-tier items in the mathematics exam. The first model attempted was a dichotomous model in which all the portions from the two-tier items were treated as if they were entirely independent items. Accordingly, all the portions were scored either as right or wrong. This model served as the baseline model in this study and was adopted purely for comparison purpose. Since it is deemed improbable by most subject matter experts that a respondent could get the second tier correct and yet missed the first tier, the second model attempted was a partial credit model in which a score of 2 was assigned to the case when both tiers were answered correctly, a score of 1 when only the first tier was correct, and a score of zero for the other combinations. This model was denoted as the 2100 model to facilitate subsequent discussion. As for the third model, another partial credit model similar to the previous one was fitted to the data with slight modification. This time, however, a score of 1 was also assigned to those respondents who obtained the correct answer to the second tier. This model was short-handed as the 2110 model below. Finally, another dichotomous model was attempted as the fourth model in which a respondent was assigned a score of 1 if and only if he/she had answered both tiers correctly. All the other combinations were scored as zero. This model was adopted upon recommendation from some subject matter experts who maintained that both the step and the answer must be correct before mastery of the content being tested could be justifiably assumed. For ease of discussion, this model was short-handed as the 1000 model. It should be noticed that since there were very few respondents who would only get the second tier correctly, hence a partial credit model with the scoring scheme of assigning a score of 3 to both tiers correct, a score of 2 to the first tier correct, a 1 to the second tier correct, and a zero to both tiers incorrect would

Table 12.4 The deviances and the reliabilities for the four-item response models being attempted

Treatment of second tier item	Deviance	Reliability
Individual items	12694.13	0.793
Scored as 2100	10792.24	0.720
Scored as 2110	10798.12	0.720
Scored as 1000	10079.44	0.710

Table 12.5 The percentages and average abilities for respondents manifesting different response patterns with respect to item 7

Response category	Percentages (%)	Average ability (logits)
Both tiers incorrect	28.37	−0.767
Second tier correct but not first	0	N/A
First tier correct but not second	5.47	−0.339
Both tiers correct	66.16	0.841

create calibration problem. For similar reason, more refined scoring schemes were not practically pursued in the present study.

The results are shown in Table 12.4 above. It is found that the fourth model had the lowest deviance statistic among the four models being processed. In addition, the reliability of the fourth model was found to be the smallest even though its value was quite comparable to the other two partial credit models. Meanwhile, the drop in reliability of the 1000 model from the first model in which the two-tier items were treated as independent items was quite prominent. There are two possible explanations for the drop in reliability when we use the 1000 model. First, when items are treated as independent items when they are actually dependent, the reliability will be artificially inflated, as in the 2100 model. Second, the three two-tier items could be more discriminating items, so a maximum score of 2 instead of 1will give more weight to these items, leading to an increase in reliability. In any case, the reliabilities among the 2100, 2110, and 1000 models are very close to each other. As a result, the 1000 model was being adopted as the model to score the two-tier items in this study.

In case further evidences were desired to justify the adoption of the 1000 scoring scheme, then more analysis should be performed at the individual item level. Reported in Table 12.5 above were the proportions of respondents who manifested different response patterns in item 7 together with their average abilities in terms of logits. As can be seen from the table, the average ability for those respondents who were incorrect in both tiers was −0.767, while that for the respondents who were incorrect in the second tier but right in the first tier was −0.339, and 0.841 for those who were correct in both tiers. It is noticed that the average abilities for those who answered both tiers correctly amounted to a positive value that was much larger than the negative average abilities for other two combinations of response categories. These findings seem to reflect that the three groups of respondents were of different abilities, with those respondents answering both tiers correctly attaining the highest average abilities while the other two groups of respondents were of closer average abilities. Thus, this further piece of information warranted strong support regarding the adoption of the 1000 scoring scheme for the two-tier items that appeared on the test, at least with respect to the models attempted.

12.6 Conclusion and Discussion

This study had demonstrated a new and rather comprehensive approach from Tam and Wu (2009) to analyze two-tier items beyond the report of mere proportions of respondents with respect to the various combinations of response categories across the two tiers as a means of data analysis. While such proportions are simple and straightforward to compute, the kind of information that can be gleaned is fairly limited. With mere proportions, incorrect responses to a two-tier item may of course be attributed to some inappropriate mastery on the part of the respondents towards the content being tested. However, it could also be attributed to some underlying deficiency in terms of the conceptualization, design, or even wordings of two-tier items being written. There is not enough information to distinguish between these and other possibilities because they are convoluted with one another. In comparison, the approach suggested herein will be much easier to comprehend by most applied researchers. Under the suggested approach, the results from the first stage of our procedure can reflect whether the two-tier item structures can be substantiated from the empirical data. With respect to the mathematics exam being analyzed, the result from the first stage will reflect whether the computational steps and their respective answers can really substantiate a two-tier structure. If there is no foundation for such claim, careful revision of the two-tier item is advised. On the other hand, the second stage aims at finding a basis concerning how the two-tier items can be scored more appropriately. The decision attained at this stage can subsequently be used to calibrate the items for various parameter estimates as well as generate other useful information.

Furthermore, it was found in this study that the 1000 model had the lowest deviance than the two partial credit models as well as the independent items model being considered. This finding forms the basis for scoring our items in accordance to the 1000 scoring scheme. Thus, consideration of an appropriate scoring procedure should constitute an important step in the analysis of two-tier items. According to our experiences, it appears that the 1000 scoring scheme always performs relatively well with two-tier items. Hence, it is suggested to always include this scoring scheme as one of the options while carrying out the second step of the suggested procedure for two-tier items.

Finally, in order to obtain the best result from the two-tier item format, it is suggested that potential item writers should try every effort to focus on improving the qualities of the items first. Pilot testing on the items is highly recommended. The procedure demonstrated in this study will be quite useful in throwing some light on the quality of the items especially during the pilot testing stage. Rather than jumping to early conclusion with regards to the abilities of the respondents, it is only after careful revision of all items with questionable quality before one should proceed to use the two-tier items for actual assessment purpose.

References

Adams, R. J., & Wu, M. L. (2011). The construction and implementation of user-defined fit tests for use with marginal maximum likelihood estimation and generalized item response models. In N. J. S. Brown, B. Duckor, K. Draney, & M. Wilson (Eds.), *Advances in Rasch measurement* (Vol. 2). Maple Grove: JAM Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum Associates.

Tam, H. P., & Li, L. A. (2007). Sampling and data collection procedures for the National Science Concept Learning Study. *International Journal of Science Education, 29*(4), 405–420.

Tam, H. P., & Wu, M. (2009). *Analyzing two-tier items with user-defined fit statistics*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, USA.

Tan, K. D., & Treagust, D. (1999). Evaluating students' understanding of chemical bonding. *School Science Review, 81*(294), 75–83.

Treagust, D. (1988). Development and use of diagnostic test to evaluate students' misconceptions in science. *International Journal of Science Education, 10*(2), 159–169.

Treagust, D. F., & Smith, C. L. (1989). Secondary students' understanding of gravity and the motion of planets. *School Science and Mathematics, 89*(5), 380–391.

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

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest – Generalised item response modeling software*. Melbourne: Australian Council for Educational Research.