Fulmer *et al. Asia-Pacific Science Education* (2015) 1:1
DOI 10.1186/s41029-015-0005-x

Asia-Pacific Science Education
a SpringerOpen Journal

RESEARCH

Open Access

Is it harder to know or to reason? Analyzing two-tier science assessment items using the Rasch measurement model

Gavin W. Fulmer[1], Hye-Eun Chu[2*], David F. Treagust[3] and Knut Neumann[4]

* Correspondence: hye-eun.chu@mq.edu.au
[2]School of Education, Macquarie University, Sydney, Australia
Full list of author information is available at the end of the article

Abstract

Two-tier multiple-choice (TTMC) items are used to assess students' knowledge of a scientific concept for tier 1 and their reasoning about this concept for tier 2. But are the knowledge and reasoning involved in these tiers really distinguishable? Are the tiers equally challenging for students? The answers to these questions influence how we use and interpret TTMC instruments. We apply the Rasch measurement model on TTMC items to see if the items are distinguishable according to different traits (represented by the tier), or according to different content sub-topics within the instrument, or to both content and tier. Two TTMC data sets are analyzed: data from Singapore and Korea on the Light Propagation Diagnostic Instrument (LPDI), data from the United States on the Classroom Test of Scientific Reasoning (CTSR). Findings for LPDI show that tier-2 reasoning items are more difficult than tier-1 knowledge items, across content sub-topics. Findings for CTSR do not show a consistent pattern by tier or by content sub-topic. We conclude that TTMC items cannot be assumed to have a consistent pattern of difficulty by tier—and that assessment developers and users need to consider how the tiers operate when administering TTMC items and interpreting results. Researchers must check the tiers' difficulties empirically during validation and use. Though findings from data in Asian contexts were more consistent, further study is needed to rule out differences between the LPDI and CTSR instruments.

Keywords: Science education, Two-tier items, Rasch measurement models, Optics, Scientific reasoning

Assessing student learning—of scientific concepts, practices, or habits of mind—is one of the central topics for research and development in science education. Such assessments can serve as formative or diagnostic tools for planning instruction and working with students, or as summative tools for gauging the effectiveness of our instructional practices, curriculum materials, or teacher education efforts. However, we observe an ongoing tension in science education assessment between our ability to construct conventional test items (e.g., multiple choice questions) that can be highly reliable but are perceived to be incapable of providing richer insights into students' conceptions and ways of thinking. Research addressing this includes efforts to make better sense of how students' responses to test items can be understood from a broader view of conceptual understanding, ability, or skill (Fulmer et al. 2014; Neumann et al. 2011).

One solution proposed that can address this tension are *two-tier items* (Treagust 1988). The two tiers in two-tier items act together to uncover students' understanding of core concepts because the student must choose a seemingly "factual" knowledge response for the first tier (Taber and Tan 2011), and then choose for the second tier what reasoning about the concept they used to arrive at the first-tier response. A large body of research across contexts has applied two-tier items to uncover students' understanding of scientific concepts as broad ranging as optics, scientific reasoning, and scientific knowledge integration and in various settings such as the US, UK, Korea, Singapore, and Australia (Chu and Treagust 2009; Johnson and Tymms 2011; Liu et al. 2011; Taber and Tan 2011; Tsui and Treagust 2009).

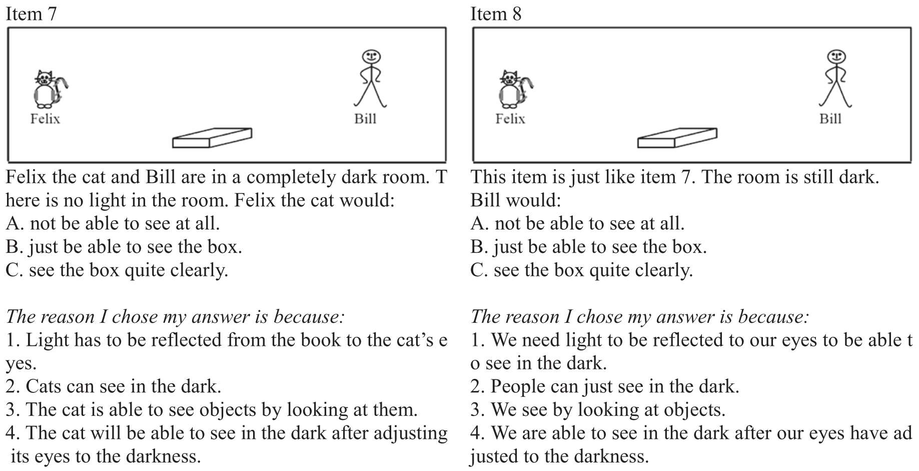
Despite the breadth of this prior research, there is still relatively little attention over how best to analyze two-tier responses and uncover how students respond to the two tiers. In particular, through rigorous measurement approaches, it is possible to examine two of the fundamental notions about two-tier items and students' responses: (1) whether the traits assessed by the first and second tiers are distinguishable yet related, and (2) whether the second tier is indeed more difficult than the first tier. For the former notion, researchers have argued that two-tier items involve related but distinct traits: for the first tier, knowing the correct answer; for the second tier, reasoning using this knowledge (Johnson and Tymms 2011; Treagust 1988). Analyzing students' responses can address this empirically. For the latter notion, prior work has posited that the second-tier portion is more difficult because it involves providing a rationale that goes beyond knowing (Taber and Tan 2011). Yet, very little research has uncovered whether students' responses on the tiers support the belief that reasoning about one's knowledge—which is indicative of understanding—is more difficult than just knowing the fact. The lack of research to answer these questions hinders the field because it threatens the validity of our use of two-tier items to examine these different cognitive skills. To address this lack of research, we conducted a study of students' responses to two-tier items to consider how the tiers are related in terms of their respective difficulty—whether identifying one's reasoning is more or less difficult than showing one's knowledge—and whether this pattern in difficulty is consistent with the distinct but related abilities the tiers are intended to assess. We do this through the application of the Rasch measurement model on both first and second tiers. This information can support ongoing research into how two-tier items uncover students' understanding and reasoning, and provide further guidance on ways to aggregate information from two-tier items into diagnostic and formative information for teachers. Thus, the purpose of the present study was to determine how first and second tiers are related.

Review of literature on two-tier items

Two-tier items have been extensively developed and researched as tools for diagnostic purposes in science education, with the seminal work by Treagust 1988 providing much impetus for the field. Tamir (1971, 1989) argued the importance of incorporating students' justifications to supplement multiple-choice test items and to evaluate students' meaningful learning experience. However, it is very time-consuming to assess students' reasoning in addition to their knowledge, especially as some students may not provide useful answers in their written statements. They often provide short written responses, which lack detail sufficient to reveal conceptual understanding and reasoning (Chu et al. 2015). To improve the ease of administration and interpretation, Two-Tier Multiple-Choice (TTMC) items incorporate a multiple-choice option for both the first and second tier, and are designed for measuring students' understanding and reasoning by providing students with opportunities to display their reasoning abilities in their justifications of the 1st tier choice (Treagust 1995). Figure 1 provides an example of a TTMC item. As Fig. 1 shows, the first tier addresses the students' knowledge of a factual point, and the second tier addresses students' reasoning. However, despite the notable progress in development and use of two-tier items, much of the work on TTMC items has focused on the diagnostic use of these instruments. There is little research that has examined the underlying relationships between the first and second tiers in terms of response patterns and students' observed difficulty in responding to these different item tiers. In the subsequent sections, we review research on two-tier items to describe their types and uses, then development and analysis of such items, and then move on to discuss a gap in the literature on two-tier items that motivates our research questions for this study.

Types and uses of two-tier items

Depending on the development framework of the TTMC diagnostic items, teachers in the classroom can assess students' understanding from various perspectives. Two-tier items have been widely used not only in Western settings but also throughout Asia. For example within the Singapore context, Chandrasegaran et al. (2007) developed TTMC items that could identify secondary school students' alternative conceptions when describing and reasoning about chemical reactions using multiple representations. Furthermore, TTMC items have been incorporated into a national project on Taiwan students' conceptual understanding (see Chiu et al. 2007, for an introduction on this national project and findings). That national project found many commonalities in results with two-tier item studies in Western contexts, but also explored how students' misconceptions could relate to specific structure and meanings of Chinese words as used in the textbooks that were unique to the setting. In another example from Singapore and with parallel study in Korea, Chu et al. (2009) developed LPDI items


Felix
Bill
Felix
Bill

Item 7

Felix the cat and Bill are in a completely dark room. T here is no light in the room. Felix the cat would:
A. not be able to see at all.
B. just be able to see the box.
C. see the box quite clearly.

The reason I chose my answer is because:
1. Light has to be reflected from the book to the cat's e yes.
2. Cats can see in the dark.
3. The cat is able to see objects by looking at them.
4. The cat will be able to see in the dark after adjusting its eyes to the darkness.

Item 8

This item is just like item 7. The room is still dark. Bill would:
A. not be able to see at all.
B. just be able to see the box.
C. see the box quite clearly.

The reason I chose my answer is because:
1. We need light to be reflected to our eyes to be able t o see in the dark.
2. People can just see in the dark.
3. We see by looking at objects.
4. We are able to see in the dark after our eyes have ad justed to the darkness.

Fig. 1 Sample two-tier multiple choice questions

specifically to assess students' stable or context-independent conceptions in two different situations: light propagation during the day and at night. The inclusion of various contexts or perspectives about the phenomena is important because, as Chu and Treagust (2014) found, only a few students were able to achieve stable, correct conceptions by perceiving the commonalities and demonstrating understanding of the scientific conceptions across contexts. Conceptual complexity and naive reasoning based on earlier experiences are particularly influential in affecting students' consistency in applying concepts. For the example of LPDI items, students who did not fully understand how "non-luminous objects reflect light" chose wrong answers and reasons for the answer choice in the given situations. So, TTMC diagnostic items require a clear purpose and an established item development framework to help teachers interpret the outcomes and plan how to draw on the material to teach the target concepts.

Two-tier items have been developed and tested for several subject areas. For example in chemistry, Tan et al. (2002) developed a diagnostic instrument on inorganic chemistry and qualitative analysis. Similarly, Taber and Tan (2010) proposed a diagnostic two-tier questionnaire on ionization energy. In biology, Tsui and Treagust (2009) developed a two-tier instrument for students' understanding about genetics. Two-tier instruments have also been developed that focus less explicitly on scientific content but on other outcomes of interest in science education. In a well-known example, Lawson (1978) developed a test of formal reasoning, later revised into the Classroom Test of Scientific Reasoning (CTSR; Lawson 2000). The CTSR has been examined in combination with tests of conceptual understanding such as the Force Concept Inventory (Bao et al. 2009; Ding 2014). Two-tier items have also been proposed that incorporate rating scale measures. For example, Bennett and Hogarth (2009) developed an instrument on attitudes towards science and school science. In their first tier, students would agree, disagree, or remain neutral about a descriptive statement (e.g., "Science lessons are among my favorite lessons"). For the second tier, the students provided a reason matching the available descriptive statement (e.g., "I like the parts about physics"). This instrument was later used by Oliver and Venville (2011) in their case study of Olympiad students' attitudes and passion for science. One important difference for this use was that, in the second tier, respondents could choose *any* or *all* of the reasons that suited them—not only one response.

Development and analysis of two-tier items

One approach to develop TTMC items begins with identifying propositional content knowledge statements on the topic, then creating a concept map that accommodates the propositional statements, a review of specific concept-related literature, interviews with students to investigate students' conceptions and reasoning, and the design of a specification grid to ensure that the developed TTMC diagnostic instrument fairly covers the topic (Treagust 1988; Treagust 1995). Recently, classroom observations have been included as part of item development to investigate how the concept or topic was actually taught in different contexts (Chu and Treagust 2014; Chu et al. 2015). This underscores the importance of incorporating the different aspects of the topic and various contexts for the concept that is being tested.

The use of TTMC items expands on previous conceptual studies and can connect with teachers' classroom practices and use. While there is a rich literature on students' conceptions (Duit 2009), science educators have realized that teachers do not have many opportunities to read these research outcomes (Duit and Treagust 2003). Developing TTMC items draws on literature on students' alternative conceptions that have been conducted over the past 20 years or so. For example, Chu and Treagust (2009) developed the Light Propagation Diagnostic Instrument (LPDI) based on prior qualitative research studies on students' conceptions mainly about fundamental geometric optics (Andersson and Kärrqvist, 1983; Fetherstonaugh and Treagust 1992; Galili and Hazan 2000; La Rosa et al. 1984; Langley et al. 1997). Thus, using TTMC items affords teachers access to results of this literature for uncovering students' pre-instructional understanding and reasoning about the phenomena or concepts.

Science teaching should facilitate students' conceptual understanding of scientifically accepted concepts and develop their capacity to reason scientifically. Part of achieving this goal is played by effective assessment procedures that are well matched with constructivist classroom teaching approaches (Bell and Cowie 2001; Black and Wiliam 1998; Treagust et al. 2001). Items that require students' reasons for selecting a knowledge answer, whether informally and formally, form a major part of effective assessment (Wiggins and McTighe 1998), and are the foundation of item development in TTMC instruments. Current efforts in TTMC also consider test and item function to examine the validity of the questions' results about students' knowledge and reasoning.

The analysis of two-tier items can proceed in a variety of ways. One approach is to give points for each tier, so that a question could be worth up to two points (for a correct tier 1 and a correct tier 2). A second option is to give credit only when both tiers are answered correctly. These decisions convey some meaningful assumptions and interpretations. For example, giving credit only for correct responses on both tiers conveys that being correct on tier 1 is only meaningful if the student can provide the appropriate rationale. On the other hand, scoring each tier separately assumes that it is acceptable for students to provide the correct rationale even if they select a tier-1 response that cannot match it. However, neither of these traditional approaches to handling the two-tier responses take into account the fact that some items, and the tiers themselves, may be harder. This is partially addressed in work that considers differences based on aspects of the concept or various contexts in which the concept is applied (Chu and Treagust 2014; Chu et al. 2015, Treagust 1995). However, this still does not uncover the differences for the traits represented by the tiers themselves. By contrast, a modern test theory approach such as the Rasch measurement model is able to account for differences in the difficulty of the different tiers above and beyond any role of content differences. Rasch measurement can also be used to examine open-ended second tiers, such as work by Liu et al. (2011).

Research questions addressed by the present study

In this study, we apply a Rasch measurement model to TTMC items to examine whether responses on the two tiers correspond to related but slightly different traits: tier 1 is about knowledge of the concept; tier 2 is about reasoning about the concept for the specific context. Though these cognitive skills are distinct, we cannot assume whether one trait is easier or harder. We may conjecture about expected difficulty, but this should be tested empirically. For example, suppose that we think that it can be easier for students to show their propositional knowledge or to make a choice that reflects their understanding of the context (tier 1) than it is to reason through their choice (tier 2). If so, then we would expect the second-tier item to be more difficult than the first-tier item. This is an empirical matter that can be addressed through Rasch measurement. Thus, by examining the difficulty of item tiers, we can understand whether the tiers demonstrate differences in difficulty, and by extension, if one cognitive skill is easier or harder for students. Additionally, because previous work has emphasized the importance of understanding the differences in difficulty across content aspects of the concept, we examined the items based on the content alone, and on an interaction of content with tier. This approach is taken because we do not yet know if the cognitive skills associated with the tiers exhibit a consistent pattern of difficulty across all content (for example, that tier 1 is always easier than tier 2), or whether the tiers vary in difficulty according to content. So we can also examine item difficulty by content areas and by the interaction of tier and content area. We do so by applying a sequence of four measurement models.

The first model estimated was a unidimensional model, in which all items—regardless of content or tier—were analyzed as if they were part of the same underlying construct. The first model corresponded with the assumption that all items addressed the same trait of *understanding*, without distinguishing *knowing* from *reasoning*. In the second model, we analyzed items according to tier: one dimension for tier-1 items to represent *knowing*, and another dimension for tier-2 items to represent *reasoning*. This second model examined if the items can be distinguished by the trait, but if there is no difference according to the items' content. In the third model, we analyzed the items by content only. The third model ignores any difference by tier, and considers only the role of understanding different content aspects of the concept. In the fourth model, we analyzed items by combining both content and tier—so every combination of content and tier represents its own dimension. This fourth model examined whether the items can be distinguished by *both* content of the item and by tier.

We ask the following research questions:

1. Which of the four measurement models best fits the students' responses?
2. How do the items differ in average estimated difficulties between tier-1 and tier-2?
3. How do the items differ in average estimated difficulties according to content?

Methods

Data for this article come from two separate studies, each using two-tier items within a study of students' scientific understanding or scientific reasoning. The context and instruments are described for each study separately. For both data sets, because each two-tier item consists of two items, a tier-1 item and a tier-2 item, the term *item* will refer to the tier-1 or tier-2 item only, whereas when we refer to the two-tier-item as a whole we will use the term *two-tier item*.

Data set 1 – optics

The first data set comes from a study of secondary students' understanding of fundamental concepts in optics—light propagation and visibility (Chu and Treagust 2009). The original work focused on whether and how students' understanding of the content was dependent upon the particular context of the items. For example, it is possible to ask questions about light propagation that take place in the daytime, or in the nighttime. The sample for this study comes from 2382 secondary students in Korea and

Singapore. Students in both samples completed a questionnaire with 8 two-tier items, thus comprising a total of 16 distinct items. Each item was scored dichotomously with a score of 1 for the correct answer and a score of 0 for any incorrect answer.

Data set 2 – scientific reasoning

The second data set comes from a study of 582 undergraduate students' attitudes toward science, epistemological beliefs, and scientific reasoning (Fulmer 2014 for more information on the original study). These undergraduates from a large, public university in the eastern USA completed a survey in exchange for course credit as part of an introductory psychology course. Among the questionnaires was Lawson's (2000) revised CTSR that consists of 12 two-tier items (i.e., a total of 24 items), organized into four aspects of reasoning: Control of Variables (8 items); Combinatorial Reasoning (8 items); Proportional Reasoning (4 items); and Probabilistic Reasoning (4 items).

Analyses

Our analyses focus on two aspects. First, we see if the relationships among items indicated that the tiers represent distinct traits, or whether the first and second tier items are indistinguishable. Second, we example the relative difficulty of item tiers to understand whether one cognitive skill (e.g., knowing) is easier or harder for students. For all of our data sets, the tier-1 and tier-2 multiple choice options were scored dichotomously, with correct response scored as 1, and any incorrect response scored as 0.

We analyzed the items using a sequence of Rasch measurement models as described above: (1) a unidimensional model that does not account for content or tier differences; (2) a tier model that accounts for first- and second-tier but does not distinguish content; (3) a content model that accounts for content differences in the TTMC items but does not distinguish tier; and (4) a combined model for content and tier. For the optics data set, models (3) and (4) have two content areas: one for light propagation, another for visibility. For the CTSR data set, models (3) and (4) have four content areas: control of variables, combinatorial reasoning, proportional reasoning, and probabilistic reasoning. Table 1 provides a summary of the models applied for each data set. As Table 1 shows, the optics data set involved four dimensions and the CTSR data set involved eight dimensions.

Table 1 Sequence of Rasch measurement models estimated for each data set

Model		Dimensions	
Number	Terms	Data Set 1 (LPDI)	Data Set 2 (CTSR)
1	None	Unidimensional(17 parameters)	Unidimensional(25 parameters)
2	Tier	Two dimensions:Tier 1; Tier 2(19 parameters)	Two dimensions:Tier 1;Tier 2(27 parameters)
3	Content	Two dimensions:Light Propagation; Visibility(19 parameters)	Four dimensions:Control of Variables; Combinatorial Reasoning; Proportional Reasoning; Probabilistic Reasoning(34 parameters)
4	Tier × Content	Four dimensions:Tier 1; Tier 2×Light Propagation; Visibility(26 parameters)	Eight dimension:Tier 1; Tier 2×Control of Variables; Combinatorial Reasoning; Proportional Reasoning; Probabilistic Reasoning(60 parameters)

Note. All models were estimated as dichotomous Rasch models. The Model 4 for both data sets was estimated as between-item models, meaning that an item was assigned to a combination of tier and content. The parameters listed indicate the complexity of the statistical model by representing the combination of dimensions and items in that model

In Rasch measurement, individuals' ability and items' difficulty are estimated simultaneously and can be compared on a common scale, called a logit scale, based on the chance that each individual would answer each item correctly. To allow such scaling requires fixing the scale to provide an anchor for these estimates. Typically, this scaling involves either constraining the set of items to have an average difficulty or zero, or constraining the persons to have an average ability of zero. Because our analyses were focused on the comparison of items, we constrained the cases (i.e., the students) so that the average ability estimates would be zero. This approach allows us to compare item difficulties (Adams et al. 2012), but it does also influence the interpretation of students' abilities. For example, analyses results may suggest that one dimension's items are easier but this may also mean that students are more able on this dimension. The distinction between whether the items are easy or the students are more able on a given dimension cannot be made without external criteria or judgment of some kind. Further exploration of the implications of this issue is raised in the discussion section.

For each data set, we compared the four models in two ways consistent with approaches for data-model comparisons. First, we examined the AIC (Akaike's Information Criterion), the AIC with correction (AICc), and BIC (Bayesian Information Criterion) values, which consider the model fit after accounting for the number of parameters in the model. An accepted rule of thumb is that models with lower AIC and BIC are considered better-fitting. Second, we used likelihood-ratio tests to compare differences in the overall fit for each model versus previous models, to test for statistically significant changes in the model deviance. This is similar to tests of changes in model R^2 in a typical ANOVA or regression. The best-fitting model was then selected for further analysis and comparison, including item-specific analyses. This included examining item fit statistics, scale reliabilities, and comparing difficulties between the modeled dimensions.

Results

Data set 1

Results show that model 4, the four-dimensional model (see Table 1), is superior: it has lowest AIC, AICc, and BIC values, and the likelihood-ratio tests show statistically significant improvements in model deviance compared to both of the 2-dimensional models. This means that the LPDI data reveal that the students and items can be distinguished not only by content in the items, but also by the tier of the item. Furthermore, the item fit statistics for the 4-dimensional model are all within a good, acceptable range for the mean-square fit statistic (between 0.7 and 1.3; Boone et al. 2014; Liu 2010). We do not consider the t-values for these fit statistics because there are a large number of cases (2000+ students), making it quite likely to find large t-values regardless of the quality of fit (Bond and Fox 2007). The dimensions all have moderately acceptable reliability: expected a posteriori/plausible value (EAP/PV) separation reliability indices range from 0.62 to 0.67. EAP/PV reliability indices are an estimate of how reliably the items can be used to distinguish students' underlying abilities.

Having accepted the 4-dimensional model according to overall model fit and item fit, we then consider the difficulty of the items (Table 2). As Table 2 shows, the tier-2 items are all relatively more difficult than their tier-1 counterparts. Additionally, the overall difficulties reveal that tier 1 is quite easy (average item difficulty = –0.61) compared to

Table 2 Average Rasch difficulty estimates for the LPDI items by tier and content

	Tier		
Content	Tier 1	Tier 2	*Average by Content*
Propagation	−1.04	−0.03	*−0.53*
Visibility	−0.17	−0.02	*−0.10*
Average by Tier	*−0.61*	*−0.02*	
Model Fit statistics:	AIC	AICc	BIC
	45688.68	45689.23	45838.85

Note. All measurement models were estimated as dichotomous Rasch models. The reported model fit statistics of AIC, AICc, and BIC are used to select this statistical model over others (not reported in this table). This model is the 26 parameter model

tier 2 (average item difficulty = −0.02). The fact that both have negative average difficulty estimates indicates that the items are, on average, relatively easy for this sample. Furthermore, light propagation is an easier content topic (average item difficulty = −0.53) than visibility (average item difficulty = −0.10). Again, the result that both content areas have negative average difficulty means that the full set of items is relatively easy for this sample of students, on average.

We can furthermore look at patterns of correlations among the students' estimated abilities based on the items in each dimension (Table 3). The correlations of dimensions according to content areas are higher than correlations that are according to tier. As can be seen in Table 3, the correlation of visibility-tier-1 with visibility-tier-2 is .92 (B with D in Table 3), and the correlation of propagation-tier-1 with propagation tier-2 is .72 (A with C in Table 3). Looking within tier shows lower correlations: tier-1-visibility with tier-1-propagation is .55 (A with B in Table 3), and tier-2-visibility with tier-2-propagation is .44 (C with D in Table 3). These results corroborate the importance of determining difficulty for both tiers, which supports the decision to consider both tier and content in our Rasch measurement model analyses. For example, because of the correlations by content area, a student who can answer tier-1 items about light propagation is more likely to get tier-2 questions about light propagation correct. There is a weaker correlation between answering tier-1 questions on light propagation and answering tier-1 questions on visibility.

Data set 2

Results show that model 4, the eight-dimensional model (see Table 1), is superior: it has lowest AIC, AICc, and BIC values; and likelihood-ratio tests show statistically significant improvements in model deviance compared to both of the 2-dimensional

Table 3 Correlations between Rasch student ability estimates for the LPDI items by tier and content

Dimension		(A)	(B)	(C)
Tier 1 × Light Propagation	(A)			
Tier 1 × Visibility	(B)	0.55		
Tier 2 × Light Propagation	(C)	0.72	0.42	
Tier 2 × Visibility	(D)	0.56	0.92	0.44

Note. The correlations shown are between the estimates of students' ability for the respective combination of tier and content models. This means that the CTSR data support that the students and items can be distinguished not only by content in the items but also by the tier of the item. The item fit statistics for the eight-dimensional model show that all items have good, acceptable fit (mean-square fit statistic between 0.1 and 1.3) for the information-weighted fit and only a few items have unweighted mean-square fit statistics above 1.3. The student responses for data set 2 thus have generally poorer fit to the Rasch measurement model than the student responses for data set 1. Even so, the items show moderate to good reliability for the person measures: EAP/PV reliability indices range from 0.64 to 0.72.

We next consider the difficulty of the items for the eight-dimensional model (Table 4). As the table shows, there is no clear pattern in difficulty of the items according to tier, because both tiers have average item difficulty estimates that are negative. Rather, the tiers differ in difficulty according to the content. Combinatorial Reasoning follows the pattern similar to the LPDI items in data set 1: tier 1 items are somewhat easier (average difficulty estimate of −0.16) than the tier 2 items (average difficulty estimate of 0.02). However, the content areas of Control of Variables and Probabilistic Reasoning are both relatively easy (both tiers have relatively large, negative values). Moreover, the content of Proportional Reasoning shows the opposite pattern: tier 1 items are actually harder on average (average difficulty estimate of 0.57), whereas the tier 2 items are relatively easier on average (average difficulty estimate of −0.40).

The correlations for students' estimated ability according to the CTSR dimensions (Table 5) tell a similar story to the correlations for the LPDI. The correlations within content areas are much higher than correlations within tier. For example, the correlation of control of variables between tier 1 and tier 2 is 0.99 (A with E in Table 5), and the correlation of the tier 1 and tier 2 for proportional reasoning is also 0.99 (D with H in Table 5). Likewise, when looking within tiers but across content areas, there is a lower correlation, such as between tier 1 probabilistic reasoning and tier 1 proportional reasoning (C with D in Table 5) which has a value of just 0.57. These results are consistent with the importance of accounting for item content in studying the items' difficulty, as in this case there is not an overall pattern by tier.

Table 4 Average Rasch difficulty estimates for the CTSR items by tier and content

	Tier		
Content	Tier 1	Tier 2	*Average by Content*
Control of Variables	−1.23	−1.22	*−1.22*
Combinatorial Reasoning	−0.16	0.02	*−0.07*
Probabilistic Reasoning	−3.08	−3.32	*−3.20*
Proportional Reasoning	0.57	−0.40	*0.08*
Average by Tier	*−0.88*	*−1.02*	
Model Fit statistics:	AIC	AICc	BIC
	12558.08	12572.13	12820.07

Note. All measurement models were estimated as dichotomous Rasch models. The reported model fit statistics of AIC, AICc, and BIC are used to select this statistical model over others (not reported in this table). This model is the 60 parameter model

Table 5 Correlations between Rasch student ability estimates for the CTSR items by tier and content

Dimension		(A)	(B)	(C)	(D)	(E)	(F)	(G)
Tier 1 * Control Of Variables	(A)							
Tier 1 * Combinatorial Reasoning	(B)	0.82						
Tier 1 * Probabilistic Reasoning	(C)	0.69	0.66					
Tier 1 * Proportional Reasoning	(D)	0.57	0.60	0.57				
Tier 2 * Control Of Variables	(E)	0.99	0.82	0.715	0.59			
Tier 2 * Combinatorial Reasoning	(F)	0.81	0.94	0.77	0.61	0.82		
Tier 2 * Probabilistic Reasoning	(G)	0.72	0.68	0.99	0.53	0.74	0.79	
Tier 2 * Proportional Reasoning	(H)	0.66	0.67	0.65	0.99	0.67	0.68	0.62

Note. The correlations shown are between the estimates of students' ability for the respective combination of tier and content

Discussion

Our purpose was to examine the relationships among items in two-tier multiple-choice (TTMC) items, to compare if there were any systematic differences in difficulty according to tier, according to content, or as a combination of these. Addressing this purpose allowed an exploration of whether existing instruments support the notion that cognitive abilities associated with the tiers are systematically easier or difficult. Our findings for the LPDI showed consistently that tier 2 items are more difficult overall, so the second tier gives more evidence of student ability. This finding was also consistent across items of different types, which supports the interpretation that the ability needed to answer second-tier items is more advanced than the ability required to answer first-tier items. That means that a student's correct response to the second-tier LPDI item exhibits more ability than a student's correct response to the first-tier. So, one option for future work with the LPDI is to give more weight to correct responses on the second tier. This weighted scoring system would give a more accurate score for students' ability by accounting for the second tier items being harder.

Our findings for the CTSR do not show the same pattern of higher difficulty for second-tier items as the LPDI. For the CTSR, there was a substantial difference based on each items' content. Some topic areas had no difference in difficulty across tiers, whereas for others there was some apparent difference. This observation underscores that a pattern of item difficulty like we saw with the LPDI *cannot* be presumed by an instrument developer or user. Thus, researchers using TTMC instruments need to check the items' difficulty to allow analysis of whether the patterns in difficulty make sense for the content and the use of the assessment. In the case of the CTSR, more research may be required to uncover the substance of students' reasoning about the second-tier items and review the validity of instrument. Regardless of the test itself, the findings also emphasize that item developers should provide information to the assessment users on response patterns and items. If item developers are expecting to use a weighting approach, then they must plan for this during the item creation stage, too—and be sure not to intentionally create questions with much harder first-tier items.

We have considered several potential explanations of the apparent difference in our findings for the LPDI and the CTSR. A first potential explanation is the content and students' understanding of it. The content of optics may have a more established basis for developing tiered items. On the other hand, the content of scientific reasoning may be markedly more complex, making the difficulty of item tiers harder to predict or control. A second potential explanation is instrument construction. The LPDI is more recently constructed than the original CTSR and draws on much more literature related to two-tier diagnostic item development. So, the item development framework for the LPDI's second tier items may be better tied to research on student thinking about optics. Lastly, the third potential explanation is in sample differences. On one hand, our study combines data from secondary students and from university students. Yet the LPDI was created with secondary students in mind, and the CTSR was created with university students in mind, so this sample problem may be reduced. On the other hand, our data come from settings in East Asia (for LPDI only) and the United States (for CTSR only). Given these unbalanced samples, there may be difference in the populations in the combined data sets that we cannot control or remove with the current data.

Limitations

Though the study yields insights about TTMC items, there are some limitations to recognize. One potential limitation is the somewhat low reliability coefficients for each of the estimated four dimensions. Using the EAP/PV separation reliability statistic, the reliability indices for the 4 dimensions in the LPDI are between 0.62 and 0.67, and the indices for the eight dimensions of the CTSR are between 0.64 and 0.72. These are all considered "moderately acceptable" for reliability. The finding is actually surprisingly good, given that there are only 4 items for each of the dimensions of the LPDI, and between 2 and 4 items for each dimension of the CTSR. The fact that they are in the moderately acceptable range is promising considering the low number of combined items and tiers (Boone et al. 2014). This finding indicates that the items are acceptable for diagnostic purposes. Yet the reliability indices are not sufficiently high for these items to be used for any high-stakes purposes, which require indices in the range of 0.80 and above (DeVellis 2012). Thus, further research on tier and content effects on item difficulty may require tests with more items for each dimension to be analyzed if the responses are intended to be scored for grading or achievement purposes.

A second limitation is based on the estimation process used. As mentioned in the Methods section, our data were constrained by cases to allow better comparisons of the items' difficulty. However, this choice coerces the students' ability estimates on the two dimensions so that both have an average of zero. This ignores the fact that students may actually be more able on one dimension—in this case, tier 1. While this does not change the pattern of the findings that we report, it does mean than any assertion about tier 1 being easier may alternatively be interpreted to mean that students' are more able on this dimension. This potential confound in interpretation is an artifact of having tier 1 and tier 2 be assigned to different dimensions. This limitation stems in part from how we conduct our Rasch analysis—because we assign each item to only one trait. This extends from our assumptions and decisions; it is not necessarily a limitation for all Rasch analyses. One potential solution would be to assign *both* tiers to a dimension representing "knowledge" and tier 2 to a second dimension for "reasoning."

This could be an interesting avenue for further study that would anchor the items across dimensions. However, all approaches that involve Rasch analysis will have to cope with the fact that the approach always attends both to items' difficulties and to students' abilities. Additionally, because our two data sets do not overlap based on instrument or on sample, we cannot compare the students across data sets on their ability, and we cannot compare the CTSR and the LPDI on their difficulty.

Implications for future research

Our findings have potential implications for how we construct two-tier multiple-choice assessments and what we do with the results of the analyses. Because we have shown that researchers cannot assume that two-tier items follow any strict pattern in terms of difficulty, subsequent work to develop and validate two-tier items needs to include efforts to gauge relative difficulty. This general exhortation to assessment developers is part of an extended push to encourage instrument developers and users to understand that raw student response data is not based on a typical ratio scale (Boone et al. 2014)—so that simple summation of test scores may be inappropriate in some circumstances. For two-tier items, there are meaningful differences in possible interpretations for the *content choice* and *reasoning* tiers, which makes the summation of scores particularly inappropriate for such instruments. Assessment developers who create two-tier instruments may wish to report not only information on the overall performance of students, but also information on how the items may have varying difficulty according to content and to tier. This recommendation holds whether the developers use Rasch measurement or use simpler statistics such as student response patterns or item facility indices.

Within Asian contexts, a growing body of research has explored the use of two-tier items, some entirely multiple choice and others with open-ended components (Taber and Tan 2011; Tan et al. 2002; Tsui and Treagust 2009). In the current findings, the data from the LPDI, which were collected in Asian settings (i.e., Singapore and Korea) showed results consistent with the notion that tier-2 reasoning items would be harder than tier-1 knowledge items. However, our findings emphasize that one cannot assume that two-tier items function as predicted without empirical validation. Though not all researchers can apply the Rasch model, other approaches to comparing the first and second tiers can be used. Furthermore, because the data sets differ by country and age, without overlap in the instruments, we cannot say that any differences for LPDI are due entirely to the Asian context, and requires further study.

Given that our data sources differ by country as well as age—and particularly with different instruments for the two data sets—several directions for further study could help us understand findings on item difficulty are replicable across contexts, ages, etc. For example, subsequent work in Korea or Singapore to find a matching university sample for the CTSR may allow comparisons with the US university data set, or a US secondary sample could be used to compare the LPDI with the current samples from Korea and Singapore. Similarly, future research could compare across educational levels: for example, including secondary students for CTSR or administering the LPDI in tertiary settings. Such studies can provide more balanced samples that can help distinguish whether the present findings are sample-dependent or are consistent for these instruments across samples.

Finally, another implication for research is to extend the CTSR by considering multiple contexts in which scientific reasoning is exhibited. This study sits within a broader trend of research to understand whether performance results match hypothesized dimensions or levels of ability (Fulmer et al. 2014; Neumann et al. 2011), and whether item features and contexts influence how students respond to the questions (Nehm and Ha 2011). Consistent with this, the LPDI intentionally considered a variety of contexts in which light propagates and visible objects could be applied. But the CTSR does not appear to be constructed in a way to address varying contexts so purposefully. Indeed, many of the CTSR items use contexts only from biological and life sciences. A fruitful direction for the field could be to consider what conceptual understanding of scientific reasoning can look like across contexts. This may mean varieties of everyday contexts as well as using different science disciplines to inform how to craft two-tier items on scientific reasoning. Through this work, revised instruments could intentionally use item context as a consideration in measurement: either to draw upon a variety of contexts to measure reasoning in various ways; or to strategically select only one or two contexts to control for variations, like the LPDI does.

About the Authors

1. Gavin W. Fulmer (Email: gavin.fulmer@nie.edu.sg): Gavin is an Assistant Professor of Curriculum, Teaching, and Learning at the National Institute of Education (NIE), Nanyang Technical University, Singapore. His areas of research are the assessment of students' science knowledge and teachers' classroom instruction, and the alignment of assessments with curriculum. Prior to joining NIE, Gavin served as an Associate Program Director in the Division of Research on Learning of the US National Science Foundation (NSF). Gavin previously worked for Westat, an employee-owned contract research firm in Rockville, MD, USA. He received his B.S. in mathematics and physics from the University of Redlands and a Ph.D. in Science Education from the State University of New York at Buffalo.

2. Hye-Eun Chu (Email: hye-eun.chu@mq.edu.au): Hye-Eun is a Lecturer, School of Education, Macquarie University, Sydney, Australia. Her research interests lie in the areas of conceptual development in science learners, interdisciplinary approaches in science (e.g. environmental education, STEM education, and multicultural education), affective factors in the science classroom, evaluation of pedagogical approaches, and diagnostic/formative assessment. Prior to Joining Macquarie University, She had been working as an assistant professor at the National Institute of Education (NIE), Nanyang Technological University in Singapore. Before Hye-Eun joining in NIE, she spent 3 years as a postdoctoral research fellow working with Professor David Treagust in the Science and Mathematics Education Centre (SMEC) of Curtin University of Technology in Perth and was funded by research grants from Dankook University in Korea and the Korean Research Foundation. She is serving as a co-editor of the *Asia-Pacific Science Education Journal* and editorial board member of the *International Journal of Science and Mathematics Education*.

3. David F. Treagust (Email: D.Treagust@curtin.edu.au): David is a John Curtin Distinguished Professor in Curtin University. David began a career as a science educator in 1964 when he began to teach high school science in Bradford England. In 1966 he came to Australia where he taught science in high schools and secondary colleges in Tasmania and Western Australia. In 1974 he commenced his postgraduate studies in science education at the University of Iowa. After graduating with a doctorate in science education, David joined as a postdoctoral fellow at Michigan State University. Then, David joined the Science and Mathematics Education Centre at the Western Australian Institute of Technology (now Curtin University). He served as doctoral advisor for a large number of students. His main research areas are conceptual change, multiple representations and chemistry education. He was President of NARST from 1999 to 2001 and served in a similar role for the Australasian Science Education Research Association from 2003 to 2010. David was also an advisor to the PISA International Science Assessment Project. He is an author of books such as Improving Teaching and Learning in Science and mathematics. He has exercised editorial leadership for international journals such as: International Journal of Science Education, International Journal of Science and Mathematics Education, Science Education, Research in Science Education, Journal of Biology Education, Australian Science Teachers Journal and Research in Science and Technological Education.

4. Knut Neumann (Email: neumann@ipn.uni-kiel.de): Knut is an Associate Professor at Leibniz Institute for Science and Mathematics Education (IPN), Germany. He is currently working on project such as: Development of physics competency in high school (DFG), Quality of Instruction in Physics (BMBF), and Mathematics and Science for life (MaScil) (EU). Knut was a research scientist at University of Düsseldorf, and received his Ph.D. in physics education from the University of Education at Heidelberg. He joined research group and graduate school "Teaching and Learning of Science" at University Duisburg-Essen as a research scientist from 2004 to 2008. Prior to joining IPN, Knut worked as

managing director of the Centre of Empirical Research at the University Duisburg-Essen. He is serving as an associate editor for Journal of Research in Science Teaching.

Author details

[1]National Institute of Education, Nanyang Technological University, Singapore. [2]School of Education, Macquarie University, Sydney, Australia. [3]Science & Mathematics Education Centre, Curtin University, Perth, Australia. [4]Leibniz Institute for Science and Mathematics Education (IPN), University of Kiel, Kiel, Germany.

Received: 5 August 2015 Accepted: 2 October 2015
Published online: 21 December 2015

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2012). *Reference manual for ACER ConQuest 3*. Camberwell, Australia: ACER.

Andersson, B., & Kärrqvist, C. (1983). How Swedish pupils, aged 12–15 years, understand light and its properties. *European Journal of Science Education, 5*(4), 387–402. doi:10.1080/0140528830050403

Bao, L., Cai, T., Koenig, K., Fang, K., Han, J., Wang, J., et al. (2009). Learning and scientific reasoning. *Science, 323*, 586–7.

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education, 85*(5), 536–553.

Bennett, J., & Hogarth, S. (2009). Would you want to talk to a scientist at a party? High school students' attitudes to school science and to science. *International Journal of Science Education, 31*, 1975–98.

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

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (2nd ed.). Mahwah: Lawrence Erlbaum Associates.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht, the Netherlands: Springer.

Chandrasegaran, A. L., Treagust, D. F., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical using multiple levels of representation. *Chemical Education Research and Practice, 8*, 293–307.

Chiu, M.-H., Guo, C.-J., & Treagust, D. F. (2007). Assessing students' conceptual understanding in science: An introduction about a national project in Taiwan. *International Journal of Science Education, 29*, 379–90.

Chu, H.-E., & Treagust, D. F. (2009). A stratified study of students' understanding of basic optics concepts in different contexts using two-tier multiple-choice items. *Research in Science & Technological Education, 27*, 253–65.

Chu, H.-E., & Treagust, D. F. (2014). Secondary students' stable and unstable optics conceptions using contextualized questions. *Journal of Science Education and Technology, 23*(2), 238–51. doi:10.1007/s10956-013-9472-6.

Chu, H.-E., Treagust, D. F, Lim, G. S. E., & Chandrasegaran, A. L. (2015). *Efficacy of multiple choice items: Do two-tier multiple-choice diagnostic items have the power to measure students' conceptions similar to open-ended items?* Paper presented at the 11th European Science Education Research Association (ESERA) conference, Helsinki, Finland.

DeVellis, R. F. (2012). *Scale development: Theory and Applications* (3rd ed.). Thousand Oaks: Sage.

Ding, L. (2014). Verification of causal influences of reasoning skills and epistemology on physics conceptual learning. *Physical Review Special Topics - Physics Education Research, 10*(2), 023101.

Duit, R. (2009). *Bibliography: Students' alternative frameworks and science education (IPN)*.

Duit, R., & Treagust, D. F. (2003). Conceptual change: a powerful framework for improving science teaching and learning. *International Journal of Science Education, 25*, 671–88.

Fetherstonaugh, T., & Treagust, D. F. (1992). Students' understanding of light and its properties: teaching to engender conceptual change. *Science Education, 76*, 653–72.

Fulmer, G. W. (2014). Undergraduates' attitudes toward science and their epistemological beliefs: Positive effects of certainty and authority beliefs. *Journal of Science Education and Technology, 23*, 198–206. doi:10.1007/s10956-013-9463-7.

Fulmer, G. W., Liang, L. L., & Liu, X. (2014). Applying a force and motion learning progression over an extended time span using the Force Concept Inventory. *International Journal of Science Education, 36*, 2918–36. doi:10.1080/09500693.2014.939120.

Galili, I., & Hazan, A. (2000). The influence of an historically oriented course on students' content knowledge in optics evaluated by means of facets-schemes analysis. *American Journal of Physics, 68*, S3–15.

Johnson, P., & Tymms, P. (2011). The emergence of a learning progression in middle school chemistry. *Journal of Research in Science Teaching, 48*, 849–77. doi:10.1002/tea.20433.

La Rosa, C., Mayer, M., Patrizi, P., & Vicentini-Missoni, M. (1984). A stratified study of students' understanding of basic optics concepts in different contexts using two-tier multiple-choice items. *Research in Science & Technological Education, 27*, 253–65.

Langley, D., Ronen, M., & Eylon, B. S. (1997). Light propagation and visual patterns: Preinstruction learners' conceptions. *Journal of Research in Science Teaching, 34*, 399–424.

Lawson, A. E. (1978). Development and validation of the classroom test of formal reasoning. *Journal of Research in Science Teaching, 15*, 11–24.

Lawson, A. E. (2000). *Classroom Test of Scientific Reasoning: Multiple choice version*. Arizona State University: Author.

Liu, X. (2010). *Using and developing measurement instruments in science education*. Charlotte: Information Age Publishing.

Liu, O. L., Lee, H.-S., & Linn, M. C. (2011). Measuring knowledge integration: Validation of four-year assessments. *Journal of Research in Science Teaching, 48*, 1079–107. doi:10.1002/tea.20441.

Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching, 48*, 237–56.

Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a Nature of Science test. *International Journal of Science Education, 33*, 1373–405. doi:10.1080/09500693.2010.511297.

Oliver, M., & Venville, G. (2011). An exploratory case study of Olympiad students' attitudes towards and passion for science. *International Journal of Science Education, 33*, 2295–322. doi:10.1080/09500693.2010.550654.

Taber, K. S., & Tan, K. C. D. (2011). The insidious nature of 'hard-core' alternative conceptions: Implications for the constructivist research programme of patterns in high school students' and pre-service teachers' thinking about ionisation energy. *International Journal of Science Education, 33*, 259–97. doi:10.1080/09500691003709880.

Tamir, P. (1971). An alternative approach to the Construction of multiple choice test items. *Journal of Biological Education, 5*, 305–307.

Tamir, P. (1989). Some issues related to the use of justifications to multiple-choice answers. *Journal of Biological Education, 23*, 285–92.

Tan, K. C. D., Goh, N. K., Chia, L. S., & Treagust, D. F. (2002). Development and application of a two-tier multiple choice diagnostic instrument to assess high school students' understanding of inorganic chemistry qualitative analysis. *Journal of Research in Science Teaching, 39*, 283–301. doi:10.1002/tea.10023.

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education, 10*, 159–69. doi:10.1080/0950069880100204.

Treagust, D. F. (1995). Diagnostic assessment of students' science concepts. In S. Glynn & R. Duit (Eds.), *Learning science in the schools: Research reforming*. Mahwah: Lawrence Erlbaum Associates.

Treagust, D. F., Jacobowitz, R., Gallagher, J. L., & Parker, J. (2001). Using assessment as a guide in teaching for understanding: A case study of a middle school science class learning about sound. *Science Education, 85*, 137–57.

Tsui, C.-Y., & Treagust, D. (2009). Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. *International Journal of Science Education, 32*, 1073–98. doi:10.1080/09500690902951429.

Wiggins, G., & McTighe, J. (1998). *Understanding by design*. Alexandria: Association for Supervision and Curriculum Development [ASCD].