International Journal of Science Education

ISSN: 0950-0693 (Print) 1464-5289 (Online) Journal homepage: http://www.tandfonline.com/loi/tsed20

Development and Validation of a Multimedia-based Assessment of Scientific Inquiry Abilities

Che-Yu Kuo, Hsin-Kai Wu, Tsung-Hau Jen & Ying-Shao Hsu

To cite this article: Che-Yu Kuo, Hsin-Kai Wu, Tsung-Hau Jen & Ying-Shao Hsu (2015) Development and Validation of a Multimedia-based Assessment of Scientific Inquiry Abilities, International Journal of Science Education, 37:14, 2326-2357, DOI: 10.1080/09500693.2015.1078521

To link to this article: http://dx.doi.org/10.1080/09500693.2015.1078521

International Journal of Science Education, 2015
Vol. 37, No. 14, 2326–2357, http://dx.doi.org/10.1080/09500693.2015.1078521

Development and Validation of a Multimedia-based Assessment of Scientific Inquiry Abilities

Che-Yu Kuo[a], Hsin-Kai Wu[a*], Tsung-Hau Jen[b] and Ying-Shao Hsu[a]

[a]*Graduate Institute of Science Education, National Taiwan Normal University, Taipei, Taiwan;* [b]*Center of Science Education, National Taiwan Normal University, Taipei, Taiwan*

The potential of computer-based assessments for capturing complex learning outcomes has been discussed; however, relatively little is understood about how to leverage such potential for summative and accountability purposes. The aim of this study is to develop and validate a multimedia-based assessment of scientific inquiry abilities (MASIA) to cover a more comprehensive construct of inquiry abilities and target secondary school students in different grades while this potential is leveraged. We implemented five steps derived from the construct modeling approach to design MASIA. During the implementation, multiple sources of evidence were collected in the steps of pilot testing and Rasch modeling to support the validity of MASIA. Particularly, through the participation of 1,066 8th and 11th graders, MASIA showed satisfactory psychometric properties to discriminate students with different levels of inquiry abilities in 101 items in 29 tasks when Rasch models were applied. Additionally, the Wright map indicated that MASIA offered accurate information about students' inquiry abilities because of the comparability of the distributions of student abilities and item difficulties. The analysis results also suggested that MASIA offered precise measures of inquiry abilities when the components (questioning, experimenting, analyzing, and explaining) were regarded as a coherent construct. Finally, the increased mean difficulty thresholds of item responses along with three performance levels across all sub-abilities supported the alignment between our scoring rubrics and our inquiry framework. Together with other sources of validity in the pilot testing, the results offered evidence to support the validity of MASIA.

Keywords: *Multimedia-based Assessment; Computer-based Assessment; Inquiry Ability; Validation; Item Response Theory*

*Corresponding author. Graduate Institute of Science Education, National Taiwan Normal University, PO Box 97-27, Taipei 11699, Taiwan. Email: hkwu@ntnu.edu.tw

Introduction

Inquiry is authentic in learning science given its similarity to what scientists do to understand the natural world. Research has also evidenced that learning through inquiry fosters the development of students' knowledge and understandings of science (e.g. Frederiksen & White, 1998). However, doing scientific inquiry requires not merely skills but also the cognitive abilities to integrate those skills with science knowledge. Moreover, inquiry abilities have been continuously emphasized in science education since they are essential competencies to productively participate in the scientifically literate society (Ministry of Education, 1999, 2008; National Research Council [NRC], 2000). In order to better support scientific inquiry in schools, it is imperative to understand the extent to which students have developed inquiry abilities, including the abilities to pose scientific questions, plan and conduct experiments, analyze data, and generate evidence-based explanations (NRC, 2000).

Evaluating complex learning outcomes, such as inquiry abilities, usually requires performance assessments, but the development and implementation of reliable and valid performance assessments are both challenging and time-consuming (Ruiz-Primo & Shavelson, 1996). Most computer-based science assessments have thus heavily relied on multiple choice and short answer questions that are common in paper-based testing (Quellmalz & Pellegrino, 2009). Yet, the static modality and close-ended responses in the traditional paper-based testing limit measurements to narrow competences such as recognizing discrete science facts and deploying individual process skills, which possibly differ from authentic inquiry activities in science (Garden, 1999; NRC, 2001). Recent efforts have been made to leverage the capacities of computer technology to assess the complex learning outcomes of scientific inquiry (e.g. Bennett, Persky, Weiss, & Jenkins, 2010; Gobert, Sao Pedro, Raziuddin, & Baker, 2013; Quellmalz, Timms, Silberglitt, & Buckley, 2012). Such computer-based assessments (CBAs) have employed multimedia to impose dynamic, complex science phenomena and allow active involvement and interactions with simulations. These CBAs thus gathered both the processes and products of students' engagement in inquiry tasks as students' inquiry abilities. However, few of these assessments have focused on how to materialize the potential of CBAs to serve summative, accountability purposes for students of multiple grades.

This study aimed to develop and validate a multimedia-based assessment of scientific inquiry abilities (MASIA) that was designed to offer comprehensive and continuous information about students' scientific inquiry abilities. Our assessment was comprehensive to the extent that it was designed to cover the important components of the inquiry ability construct. The comprehensiveness enabled the potential to serve summative and accountability purposes of large-scale assessments because of better generalizable assessment results. Even when each student was only allowed to take a few tasks under the common time constraint of large-scale assessments, the generalizability could be extended to different components of the inquiry ability construct and different science topics beyond those few tasks (Ruiz-Primo & Shavelson, 1996).

Furthermore, the continuous information in our assessment was achieved by designing the construct with increased levels of performance complexity along with students' learning trajectories. The continuity thus extended the potentials to understanding of students' learning progress (Duschl, Schweingruber, & Shouse, 2007). The comprehensive and continuous results of our assessment will match the interests of policymakers who mandate accountability to ensure continuous progress towards academic proficiency under the current standard-based reforms (Crundwell, 2005).

Additionally, by taking advantage of computer technology, MASIA captures students' inquiry abilities that are difficult to assess in paper-based testing. We developed a collection of multimedia, scenario-based tasks that covered the most important components of inquiry abilities in our assessment construct, that were contextualized in different content areas, and that were suitable for students of different grades across the whole of secondary school in order to achieve the aimed potential. In addition to more authentically measuring inquiry abilities, our assessment will facilitate the work of both researchers and practitioners. Classroom teachers can select, link, integrate, and sequence our tasks for their teaching and assessment practices. The orchestrated tasks are helpful for informing instructional decisions for the teachers and identifying the needs of their students towards becoming proficient in inquiry learning. Through technology applications, the assessment designers can extend our assessment to individualized and/or curriculum-embedded testing (Kuo & Wu, 2013). The extended assessments will alleviate teachers' burden in the classroom while scaffolding the continuous progression of inquiry learning in a comprehensive manner.

To achieve the aim of this study and materialize the potential benefits, two research questions were explored: (1) What design steps could be taken to develop a multimedia-based assessment in order to offer comprehensive and continuous information about students' scientific inquiry abilities? (2) What evidence could be used to establish the validity of a multimedia-based scientific inquiry assessment?

In the next sections, we provide a definition of inquiry abilities for the assessment, followed by a review of studies concerning assessments of inquiry abilities and assessment validity. The possible advantages and limitations of current computer-based inquiry assessments are discussed to illuminate the need to develop and validate a CBA to measure students' inquiry abilities in a comprehensive and continuous manner for summative and accountability purposes. The rationales for the types of validity evidence we collected are also discussed. We then present the development and validation of MASIA to offer possible answers to the research questions.

Literature Review

Definition of Inquiry Abilities

Inquiry in science education refers to activities that resemble scientists' work in an effort to extend our knowledge of the natural world, and the multifaceted nature of these activities is consensually recognized (Duschl et al., 2007; Krajcik & Czerniak, 2007; NRC, 2000). The activities usually include

> making observations; proposing questions; examining books and other sources of information to see what is already known; planning investigations; reviewing what is already known in light of experimental evidence; using tools to gather, analyze, and interpret data; proposing answers, explanations, and predictions; and communicating the results. (NRC, 1996, p. 23)

Instead of passively receiving factual knowledge, engaging students in inquiry-related activities could facilitate an active way of learning science; that is, students make use of their inquiry skills and current knowledge to propose their own explanations and construct their own knowledge in relation to data and evidence. NRC (2000) thus argued that successful engagement in inquiry-related learning activities requires some fundamental abilities that integrate inquiry skills and science knowledge together. In accordance with the integrated view of inquiry, we defined inquiry abilities as the proficiencies to coordinate science knowledge and skills in the inquiry-related activities mentioned above, and laid out a framework of inquiry abilities based on this definition of inquiry abilities for designing our assessment.

Assessments of Scientific Inquiry Abilities

To authentically capture students' inquiry abilities, researchers have designed performance assessments to request hands-on demonstrations from students (Pine et al., 2006; Zachos, Hick, Doane, & Sargent, 2000). For example, Zachos et al. offered students laboratory equipment and estimated various aspects of inquiry abilities based on their direct investigations into related science phenomena with the equipment. However, research has shown that the reliability and validity of performance assessments tend to be low when evaluations are derived only from a few tasks, and when the training and monitoring of scoring are not appropriate (Ruiz-Primo & Shavelson, 1996). The limitations of performance assessments set challenges for large-scale implementation.

Given the challenges and demands from performance assessments, it is not surprising that traditional paper-based testing is still widely used. Typical inquiry items consist of item stems that describe a scenario-based problem, followed by one or several questions that prompt solutions from students regarding certain aspects of inquiry abilities (Kind, 2013; Lorch et al., 2010; Songer, Lee, & McDonald, 2003). A collection of such items across a full coverage of the inquiry ability construct has been deemed as an effective tool for tapping inquiry abilities in a comprehensive way because of its high reliability and objective scoring (Wenning, 2007). However, the static modality and close-ended responses only allow for the demonstration of skills such as recognizing discrete pieces of science knowledge, possibly differing from inquiry activities in science (NRC, 2001).

In addition to improving the efficiency of test administration and data handling, recently, CBAs have leveraged the potential of advanced technology to measure complex learning outcomes, such as inquiry abilities, over and beyond what can be assessed in paper-based testing. For example, the Organization for Economic Co-operation and Development (2010) suggested that CBAs could visually present

complex and dynamic systems of science phenomena that are invisible or inaccessible in everyday situations. Several research teams further made use of interactive and multimedia-based simulations to assess students' inquiry abilities (Bennett et al., 2010; Buckley, Gobert, Horwitz, & O'Dwyer, 2010; Gobert et al., 2013; Quellmalz et al., 2012). For instance, one task in Quellmalz et al.'s SimScientists assessment employed dynamic animations to present realistic ecosystems, and allowed students to design and conduct iterative experiments and observe, predict, and explain the emergent behaviors of the systems from their experiments. In addition to using written answers as evidence of students' inquiry abilities, the assessment system collected and analyzed data from students' interactions with the simulations.

However, there are some limitations when these simulation-based assessments are employed for summative purposes in large-scale implementation. Firstly, some assessments only employed a few tasks and focused on certain topics only. For instance, BioLogica™ reported in the study of Buckley et al. (2010) only included one simulation in each of three focused topics (e.g. genetic transmission and trait inheritance). Also, the simulation-based inquiry tasks developed by Bennett et al. (2010) were only related to physics (balloon science specifically). The generalization of these assessments may not be valid beyond the employed tasks. Secondly, students in most of these assessments conducted investigations to achieve the goals that were set by the tasks. Without offering opportunities to propose testable questions from observation of natural phenomena, the assessments may miss important components of inquiry abilities such as asking and proposing questions (NRC, 2000). The assessments thus may not measure inquiry abilities in a comprehensive manner. Thirdly, some of these assessments were designed for one grade or for a narrow grade range of students. For example, the tasks in the SimScientist assessment (Quellmalz et al., 2012) were designed for 8th graders, and the microworld tasks in Gobert et al.'s (2013) study focused on middle school students even though multiple tasks were developed in these two assessment systems. The results of the assessments may not be informative enough to understand student learning over time, and may not support the recent efforts of developing meaningful and coherent science teaching based on learning progressions (Duschl et al., 2007; Shin, Stevens, & Krajcik, 2010). To address the aforementioned limitations, we developed MASIA to serve as a large-scale assessment that covers a more comprehensive construct of inquiry abilities and which can be used for measuring learning progress across multiple years at the secondary school level.

Assessment Validity

Assessment validity refers to an overall evaluation of the appropriateness and adequacy of the assessment purposes that are typically related to the use and interpretation of test scores (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Following Messick's (1995) unified conception of construct validity for comprehensive evaluations of assessment validity, we considered six aspects of Messick's construct validity in the previous studies, particularly in those intended to develop and validate scientific

inquiry assessments. However, our examination excluded the consequential aspect of Messick's construct validity because our assessment is still experiencing refinements and has not yet been put into practice to address the actual consequences of the assessment use. Thus, combining research on the assessments tapping similar constructs, below we describe five aspects of construct validity and the types of evidence that are plausible to support each aspect of validity.

The content aspect. The content aspect of construct validity refers to the relevance and representativeness of the assessment to the content of the measured construct (Messick, 1995). To support the content relevance of inquiry-related assessments, researchers have used expert appraisal of the task quality, the appropriateness of the content of the measured construct, and the alignments between the developed tasks and the construct (Liu, Lee, Hofstetter, & Linn, 2008; Quellmalz et al., 2012). Furthermore, because covering the important content of the construct is in reality critical for evaluating the content representativeness, Messick (1995) suggested that the power of an assessment that characterizes and differentiates expertise related to the measured constructs can maintain such importance. The discriminant evidence gathered from groups with different levels of inquiry abilities is thus likely to support the content representativeness of an assessment of inquiry abilities.

The substantive aspect. The substantive aspect of construct validity is concerned with the consistencies between the empirical data and the theoretical predictions for the nature of the measured construct and the cognitive processes involved in the construct (Messick, 1995). Assessments tapping similar scientific inquiry constructs have used the dimensionality analysis to validate the theoretical arguments about the components of the construct and the relationships among the components (e.g. Bennett et al., 2010; Kind, 2013). The simulation-based inquiry assessment in Quellmalz et al. (2012) also gathered the empirical evidence through think-aloud protocols to uncover students' reasoning processes, while solving tasks, to confirm the mental processes that were predicted by the relevant theories.

The structural aspect. In contrast to the substantive aspect, the structural aspect of construct validity focuses on the consistencies between the scoring models and the theoretical predictions (Messick, 1995). To the best of our knowledge, no scientific inquiry assessments have offered empirical evidence to support this aspect of construct validity. We thus refer to a study that reported a knowledge integration science assessment to illuminate the kinds of evidence for this aspect of construct validity (Lee, Liu, & Linn, 2011). Lee et al. (2011) demonstrated increased difficulties of item responses as the assigned scores in the rubrics increased. Because the assigned scores were derived from the levels of knowledge integration framework, the pattern thus demonstrated an alignment between their scoring rubrics and their framework.

The generalizability aspect. The generalizability aspect of construct validity is intended to establish the boundaries of assessment results when the results are generalized beyond the administered situations, such as tasks, time, occasions, and raters; the traditional reliabilities that offer measurement errors from different situations are related to this aspect of validity (Messick, 1995). Two kinds of reliability, deemed as evidence of this aspect of construct validity, have been provided by the previous simulation-based inquiry assessments. Firstly, the inter-rater reliabilities for the open-ended responses are reported to evidence the generalizability across scorers (e.g. Bennett et al., 2010). The generalizability across raters is particularly important when the auto-scoring of such open-ended responses is used in simulations (Gobert et al., 2013). Secondly, when item response modeling is applied, the overall measurement errors of students' abilities (i.e. person separation reliability) are estimated to support this aspect of construct validity (Quellmalz et al., 2012). The person separation reliability is formulated mathematically similar to the conventional reliability index of internal consistency (i.e. KR20) and is able to inform the generalizability across tasks (Wright & Stone, 1999).

The external aspect. The external aspect of construct validity is concerned with the extent to which the empirical relationships of assessment scores with other measures are consistent with what is predicted by the theories relevant to the construct (Messick, 1995). For example, previous research collected information about the correlations with similar items and instruments such as scientific practice items from the American Association for the Advancement of Science project item database and Lawson's Classroom Test of Scientific Reasoning to support this aspect of construct validity in simulation- and multimedia-based inquiry and problem-solving assessments (Chung, de Vries, Cheak, Stevens, & Bewley, 2002; Quellmalz et al., 2012).

During our assessment development, we collected multiple sources of evidence from the aforementioned five aspects of construct validity. The power to discriminate student groups with different inquiry experiences, the expert appraisal of content relevance, students' think-aloud protocols, and the convergent relationships with a similar measure were, respectively, used to support the content, substantive, and external aspects of the unified construct validity of our assessment, as described in the sub-section of Conducting Pilot Testing. In addition, the inter-rater reliabilities in the sub-section of Scoring support the generalizability aspect of the validity. In the Result section, we present two analyses, the dimensionality and comparison of the scoring rubrics and the inquiry framework, to evaluate the substantive and structural aspects of the validity of our assessment after Rasch modeling was applied. Furthermore, by the modeling process, the person separation reliability was estimated to add evidence of the generalizability aspect of the validity. The item fits, the discrimination indexes, and the alignment between student ability and item difficulty derived from this modeling were also examined because the results from Rasch modeling were appropriate only when the assessment revealed satisfactory psychometric properties from these three pieces of information.

Development of MASIA

To answer the first research question, we modified the construct modeling approach (Brown & Wilson, 2011; Wilson, 2005) for use in our assessment design process. The approach focuses on designing, implementing, and evaluating four cornerstones through a process of four steps. These four cornerstones include a construct map, item design, outcome space and a measurement model, and the four steps consist of item development, scoring guide development, pilot testing, and model analysis.

In this study, we modified the model and transformed the original four steps into our five design steps for two reasons (see Figure 1). Firstly, construct-map development was designated as one unique step. This designation indicated our efforts to identify the important components of the inquiry ability construct and to delineate the expected performances at different complexity levels along with learning progressions toward becoming competent. Secondly, the item design step was extended to the task design. This extension demonstrated our attention paid to both the task conceptualization and the actual design of the computer interface. Bennett and Bejar (1998) argued that unexpected threats may change the constructs of interest in a CBA when the interface design is not taken into consideration during task development.

Our five steps included (1) Developing an assessment framework: analyzing the components of the inquiry ability construct and identifying the performance complexities of each component; (2) Designing tasks and items: designing the multimedia, scenario-based tasks and items to elicit inquiry performances across all components and complexity levels in the framework; (3) Developing scoring rubrics: developing an outcome space on the basis of the assessment framework to guide the scoring of each item; (4) Conducting pilot testing: gathering the preliminary validation evidence to support the theoretical basis of our construct; and (5) Applying the Rasch modeling

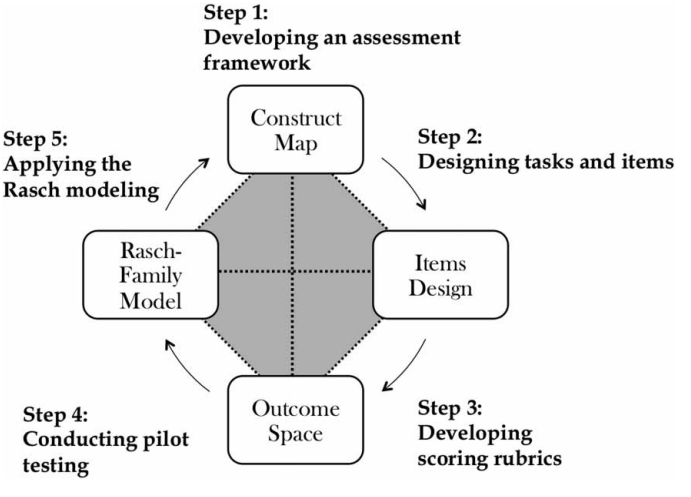

Figure 1. The five steps of assessment development modified from the construct modeling approach (Brown & Wilson, 2011)

approach: using Rasch measurement models to link the scored data back to the inquiry proficiency defined in the assessment framework.

Developing an Assessment Framework

We developed an assessment framework to set forth our assessment purposes and the precipitated constructs, in the form of an expanded construct map, to guide our design decisions in the subsequent steps. That is to say, our framework visually presented the important components of the inquiry ability construct, with each on the continuum of performance complexities, as suggested by the construct modeling approach (Wilson, 2005).

To serve the purposes of understanding the junior and senior high school students' scientific inquiry abilities in this study, we first defined our inquiry ability construct with reference to the policy reports and documents (e.g. NRC, 1996, 2000) and relevant research focusing on scientific inquiry in classrooms (e.g. Abd-El-Khalick et al., 2004; Krajcik et al., 1998). As mentioned previously, we defined inquiry abilities as proficiencies to coordinate cognitive skills and science knowledge together during students' engagement in the kind of activities similar to professional scientific discovery. Because it was not possible to exhaust all abilities involved in the multifaceted inquiry activities, after considering the nature of inquiry and the scope of the assessment, we focused on four fundamental abilities (i.e. four components of the inquiry ability construct) and related sub-abilities (Table 1) that were essential to successfully engage in the inquiry activities (NRC, 2000). These four abilities consisted of questioning (e.g. asking and identifying questions), experimenting (e.g. identifying variables and planning experimental procedures), analyzing (e.g. identifying relevant data and transforming data), and explaining (e.g. making a claim and using evidence), as suggested by the standards (NRC, 1996, 2000).

Secondly, three levels of performance complexity were identified to delineate the continuum of the inquiry ability construct for students from grades 7 to 12 rather than one (or two) specific grade(s) (Table 1). The high levels were defined on the basis of the performance expectation for students in grades 9–12 in the standards documents (NRC, 2000). The middle levels described the intermediate performances in a way with significant accomplishments demonstrated by more proficient students than novices. These levels also included typical difficulties students had in the inquiry activities according to the relevant research. The low levels were defined as the performances from novice students with little evidence of being proficient in the expected performances. Our construct-map type of assessment framework, therefore, articulated the inquiry abilities that were intended to be measured, and laid out a blueprint for a comprehensive coverage of the tasks across an appropriate range of performance complexity and across the components of the inquiry ability construct.

Thirdly, because 'inquiry abilities require students to mesh these [science] processes with scientific knowledge' (NRC, 2000, p. 18), we also incorporated the content areas into the framework as an additional dimension (Table 2). This dimension was guided by a decision to design tasks around all four areas in high school science

Table 1. Examples of three levels of performance complexity in the framework of inquiry abilities

	Inquiry abilities			
	Questioning	Experimenting	Analyzing	Explaining (causal explanation)
Sub-abilities	Q1. Formulating research questions Q2. Identifying research questions Q3. Making predictions	R1. Identifying controlled and manipulated variables R2. Planning experimental procedures R3. Selecting appropriate measurements	A1. Identifying relevant data A2. Transforming data	E1. Making a claim E2. Using evidence E3. Reasoning from evidence to the claim E4. Offering and evaluating alternative explanations
Levels of performance complexity (Using one sub-ability as an example)	Q1. Formulating research questions High: Students propose a relational or causal research question that can be answered by scientific investigations Middle: Students propose a descriptive research questions that can be answered by scientific investigations Low: Students propose an irrelevant research question or the proposed question cannot be answered by scientific investigations	R1. Identifying controlled and manipulated variables High: Students identify correct controlled and manipulated variables Middle: Students identify correct manipulated variables but the controlled variables are incorrect Low: Students identify incorrect controlled and manipulated variables	A1. Identifying relevant data High: Students recognize anomalous data and identify data relevant to research questions Middle: Students recognize anomalous data but identify data irrelevant to research questions Low: Students fail to recognize anomalous data and the data they identify are irrelevant to research questions	E1. Making a claim High: Students generate correct claims and the claims completely articulate the relationships between variables and trends of data Middle: Students generate correct claims but the claims incompletely articulate the relationships between variables and trends of data Low: Students identify or generate incorrect claims

Table 2. Item numbers of four inquiry abilities in each content area and each block

	Inquiry abilities				
	Questioning	Experimenting	Analyzing	Explaining	Total
Content area					
Physics	14	17	9	15	55
Chemistry	3	8	6	12	29
Biology	0	2	0	9	11
Earth Science	7	2	3	7	19
Total	24	29	18	43	114
Block					
A	1	7	4	9	21
B	6	3	3	13	25
C	8	9	3	5	25
D	3	7	7	6	23
E	6	3	1	10	20
Total	24	29	18	43	114

(i.e. chemistry, physics, biology, and earth science) in the next step of the task design. Two considerations led to this decision. Firstly, students' content knowledge was a possible source of irrelevant variances in their performances and might affect item difficulties beyond inquiry abilities. Secondly, incorporating the content areas into assessment tasks offered students opportunities to coordinate the knowledge of the subject areas and inquiry skills together. Thus, designing tasks across this content area dimension could balance the confounding effects of content knowledge and increase the fidelity to the integrated view of inquiry abilities.

Designing Tasks and Items

Following the blueprint laid out in the assessment framework (Tables 1 and 2), we developed scenario-based tasks with four content areas and items across four inquiry abilities. The scenario-based tasks were developed to engage students in meaningful and authentic inquiry situations. Approximately, 100 items with an equal number in each inquiry component were planned in our blueprint, given that 20 items per subscale has been recommended as providing sufficient measurement precision in large-scale assessments (Gonzalez & Rutkowski, 2010). As a result of an iterated procedure of item and task design, we developed a total of 114 items; Table 2 shows the item distribution and indicates our similar emphasis on the four inquiry abilities.

In order to materialize the potential of CBAs, we intentionally considered the applications of multimedia in our tasks, which required taking both the task and the interface designs into account. For example, a simulation was proposed to allow students to carry out systematic investigations into the camera (see an example in Figure 2). The vivid images of the camera components helped students visualize the task situation

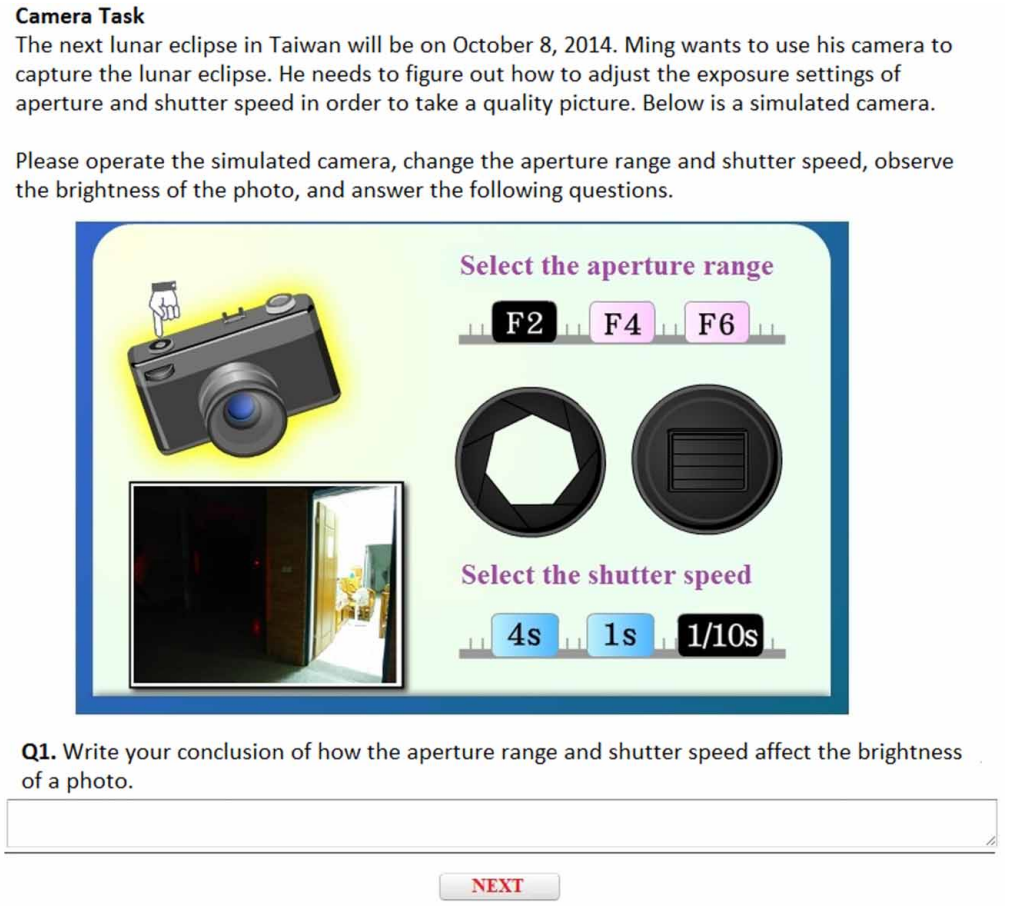

Figure 2. The camera task in MASIA. This task provided a simulated camera that allowed students to perform tasks by changing the aperture range and shutter speed in the assessment

without heavy language demands, and the linkages between their manipulations and the resulting pictures also directed their attention to the important features of the camera during their investigations. For the interface design, the features of the designed simulation were taken into careful consideration. In the camera task, we considered which features of the interface would invite students to change the size of the aperture and shutter speed in the simulation and how they would do it, how operable features would respond to their inputs, and what resulting pictures should be linked to students' input.

In order to facilitate assessment continuity, the level of performance complexity was also taken into consideration in this step. Our task and item design thus focused on populating our assessment across the three levels to allow a range of students' performances. Additionally, to effectively and efficiently distinguish different performance complexities, we employed a combination of three types of questions in our task design. The first type was specific to one of the three levels of the abilities. The

second type of questions allowed open-ended responses. The last type of questions was ordered multiple-choice questions (Briggs, Alonzo, Schwab, & Wilson, 2006) which entailed alternative options, with each choice reflecting a level of performance complexity in our framework.

Developing the Scoring Rubrics

After the tasks and items were designed, scoring guides were developed for evaluating students' responses. The development of the scoring rubrics was closely related to two cornerstones of the construct modeling approach: item design and outcome space (Figure 1). An outcome space consists of a set of qualitatively different categories for identifying, evaluating, and scoring evidence of performances in students' responses (Wilson, 2005). In this step, we drafted a content-specific outcome space for each item. For example, in the camera task A-1 (Figure 2), we created an outcome space as shown in Figure 3 and used it as the scoring rubric. The outcome space contained the level codes to indicate the performance levels in our framework and the associated performances, the score codes for actual points, the descriptions for the specific responses, and the examples of anticipated responses. In all examples, the rationales for how the response example fits with the description were included to avoid misinterpretation (in italics and in parentheses).

Additionally, in the scoring rubrics, the score points represented the difficulties of the legitimate responses in each score category relative to an alternative (or alternatives) along with two- (or three-) point scales. The assigned scoring points thus enabled differentiation among students' responses varying across levels of performance complexity.

In this step, we also applied automated scoring to exploit the potential of CBAs. However, in order to ensure the validity of the scoring, we only employed the automated scoring for selected responses. The scoring rubrics were transformed into grading algorithms to facilitate automated scoring of those responses from multiple-choice questions, pull-down selections, and marked checkboxes.

Conducting Pilot Testing

After the tasks, items, and scoring rubrics were developed, we conducted pilot testing to collect multiple sources of data for assessment validation. We first gathered the discriminant and convergent evidence from a small group of students (Wu, Wu, & Hsu, 2014). Specifically, the discriminant evidence indicated that our inquiry measures were able to differentiate between high school students who majored in the science program and in the arts and social science program as well as between those science students with and without extra-curricular experience of science fair training. This discriminant evidence suppored the content aspect of validity. In addition, the convergent evidence showed that students' performances in our inquiry tasks were significantly correlated with their scores in Lawson's Classroom Test of Scientific Reasoning. The correlation between the MASIA items and the similar instrument

Task A-1	After operating the simulated camera by changing the aperture range and shutter speed, please write down how the aperture range and shutter speed affect the brightness of a photo		
Level Code	Performance	Score Point	Response
E1.H	Students generate correct claims and the claims completely articulate the relationships between the variables and the trends of the data.	2	Description: A response includes correct claims about how the aperture range and shutter speed affect the brightness of a photo. Example: the narrower the aperture range is and the faster the shutter speed is, the darker the photo is (*including correct claims about both relationships*).
R1.M	Students generate correct claims but the claims incompletely articulate the relationships between the variables and the trends of the data.	1	Description: A response includes correct claims about how the aperture range or shutter speed affects the brightness of a photo. Example: a. When the shutter speed is kept at 4s, the wider the aperture range is, the brighter the photo is (*only including a correct claim about one relationship*). b. The faster the shutter is, the darker the picture is; the wider the aperture range is, the darker the picture is (*including one correct claim about the effect of the shutter speed but one incorrect claim about the effect of the aperture range*).
R1.L	Students identify or generate incorrect claims.	0	Description: A response includes incorrect claims about how the aperture range or the shutter speed affects the brightness of a photo or consists only of irrelevant information Example: a. The wider the aperture range is and the slower the shutter speed is, the darker the picture is (*including an incorrect claim about both effects of the shutter speed and the aperture range*). b. The wider the aperture range is and the faster the shutter speed is, the more difficulty I have taking a photo in the dark (*including a claim that is irrelevant to the brightness of a picture*). c. The wider the aperture range is and the faster the shutter speed is, the clearer the photo is (including a claim that is irrelevant to the brightness of a picture). d. Yes (*Did not answer the question*).

Figure 3. The scoring rubrics of making a claim for the camera task in MASIA. E1.H = making a claim at the high level of performance complexity; E1.M = making a claim at the middle level of performance complexity; E1.L = making a claim at the low level of performance complexity

of scientific reasoning addressed the external aspect of the unified construct validity. Together, these two sources of evidence strengthened the theoretical links of our inquiry ability construct.

Secondly, to provide further evidence of the content aspect of the validity, we collected data about content relevance based on the expert appraisal. Three professors in science education agreed that the simulation-based tasks were sufficiently relevant (kappa = 0.88–0.96) to inquiry abilities as the assessment framework indicated.

Thirdly, we analyzed the students' think-aloud protocols to illuminate the consistencies between the cognitive demands of the simulation-based tasks and the complexity of the demands of our assessment framework (Wu et al., 2014). The students' protocols showed that most of their reasoning processes were consistently characterized by the inquiry abilities as defined in our assessment framework. The consistent patterns in students' protocols thus supported the alignments between the task designs and the measured constructs in our assessment, evidencing the substantive aspect of the unified construct validity.

Finally, we carried out usability testing of MASIA. Student exit interviews were employed to reveal issues resulting from the flagged wordings of the questions, students' misunderstandings of the tasks, the unanticipated manipulations of the media and responses to the interactive tasks, unintended task demands, and unexpected cognitive processes and strategies employed by the students. We then revised MASIA accordingly in order to improve its coherence.

Applying the Rasch Modeling Approach

In the last step, we applied Rasch modeling to validate our assessment based on a large-scale implementation. This modeling allowed an alignment between student ability and item difficulty on the same scale. The alignment could facilitate an examination of the quality of the developed items and indicate a linkage between the scored data and the assessment framework. The quality of the items individually and as a whole would build up a foundation to support the power of MASIA for offering comprehensive and continuous information after the potential of providing these two kinds of information was established in the previous steps. The data-framework linkage would directly offer empirical examinations of the structures within its comprehensiveness and continuity.

In this step, we first used Rasch modeling to examine whether the individual items entailed satisfactory psychometric properties as expected by the model. Secondly, we evaluated whether the collected data supported the structure of the inquiry ability construct as four distinctive components that were specified in our framework. Thirdly, Wright maps were created to indicate the quality of our assessment as a tool to evaluate our targeted students. In the item response modeling, a Wright map is used to map both student ability and item difficulty estimates on the same scale, and to visually present such alignment (Wu & Adams, 2007). Finally, the consistencies between performance complexity levels and the item difficulties of corresponding responses were evaluated to evidence the alignment between the scoring rubrics and our inquiry framework. In the following sections, we provide a detailed account of the application of this modeling approach to offer evidence of the substantive and structural aspects of the unified construct validity with a large-scale implementation.

Validation of MASIA

To answer the second research question, a large-scale implementation with a wider range of targeted students was conducted. In this implementation, we adopted a two-stage stratified cluster sampling method to obtain a representative sample across different proficiency levels of the target population. The Balanced Incomplete Block Design (BIB) (Gonzalez & Rutkowski, 2010) was applied to divide the tasks and items of MASIA into several exclusive item groups (each group is referred to as a block) and assembled them into booklets for reducing individual students' burden under time constraint, controlling confounding effects, and obtaining accurate estimates of students' inquiry abilities. In the following Methods section, we first describe the sampling procedure and present the application of BIB to the block assembling and booklet administration. Data of six booklets were collected to derive more accurate estimates of students' inquiry abilities. Lastly, we describe the procedures of scoring and data analysis.

Methods

Sampling and Participants

For a better generalization of the results, we adopted a two-stage stratified cluster sampling method to recruit students in 8th grade and 11th grade in 16 schools in the Taipei-Keelung metropolitan areas in northern Taiwan. The two-stage stratified cluster sampling method drew a sample of schools (clusters) from different levels of proficiency (the stratified school population) in the first stage, and selected students (elements of sampled clusters) from the sampled schools in the second stage. This method was an economical way to obtain samples that covered a wide range of student proficiencies and were thereby representative of the population. This method is used in large-scale assessments such as the Trends in International Mathematics and Science Study and the Progress in International Reading Literacy Study (Martin & Mullis, 2012).

In order to implement the sampling method, 78 senior high schools and 198 junior high schools in the areas were divided into 16 strata, 8 strata at each grade level according to students' percentile ranks in the Basic Competence Test for Junior High School Students. This test result is one requirement for junior high school graduates' admission into senior high or vocational schools in Taiwan. In each stratum, we sampled one school, and within this school cluster, we invited two classes to participate in our study. This sampling method resulted in 1,068 students, 477 8th graders and 591 11th graders, participating in our study. However, two students offered no responses during the administration of the booklet. The data of these two students were therefore deleted, resulting in 1,066 students, 476 8th graders and 590 11th graders.

The Balanced Incomplete Block Design

Following the assessment development, we developed 114 items in 30 tasks for MASIA. In order to examine the psychometric properties of these items, we applied

BIB to divide these 114 items into 5 blocks and assemble the 5 blocks into 6 booklets for test administration. Table 2 presents the actual block design of our assessment.

To avoid the confounding effect of students' prior science knowledge, we first formed grade differentiated blocks. In reference to the current high school science curriculum in Taiwan, 5 researchers in the research team coded the 30 task scenarios into 3 categories: tasks for 8th graders only, tasks for both 8th graders and 11th graders, and tasks for 11th graders only. Five blocks were then formed (Table 2): 4 for 8th graders (named as A, B, C, and D) and 4 for 11th graders (named as B, C, D, and E). The four booklets of each grade were then assembled and spirally administered to students across classes and schools at the grade as suggested by the BIB design. Table 3 shows the number of students in the two grades who completed each booklet.

The BIB design was applied to the block assembly and test administration for three reasons (Gonzalez & Rutkowski, 2010). Firstly, the assessment could be administered within a reasonable amount of time. By administering booklets, students received 45–50 items to complete within an 80-minute session. Secondly, including two blocks in a booklet allowed overlapping blocks across booklets which thus provided linking items. These linking items in overlapping blocks helped scale the item difficulties and estimate the student proficiencies. Thirdly, students may perform better or worse on the items appearing later in the test, because they could become familiar with the tasks (the practice effect) or because they were distracted, tired, or not paying attention at the end of the test (the fatigue effect). To control these confounding effects, the balanced arrangement of the booklets by placing each block once in each block position was employed (Table 3).

Scoring

Before scoring, we first examined a sample of collected responses and expanded the rubrics to fit the unanticipated responses. In order to establish the inter-rater reliability, three pairs of researchers from the research team were trained to grade all typed answers and open-ended responses in two assigned booklets and proceeded to grade 10 students for each booklet. Initial disagreements were solved by discussion

Table 3. The booklet designs and the number of students taking each booklet

				Number of students		
Subjects	Booklet	Blocks		Grade 8	Grade 11	Total
Group 1	1	A	B	119	0	119
Group 2	2	B	C	119	142	261
Group 3	3	C	D	118	150	268
Group 4	4	D	A	120	0	120
Group 5	5	D	E	0	148	148
Group 6	6	E	B	0	150	150
Total				476	590	1066

of the definitions and examples of the responses. The pairs of researchers then independently graded 50 students in each booklet. The averaged percentages of the initial agreements ranged from 77% to 98%. Among these 73 graded responses, the agreements of 36 responses did not exceed 85%. The graders then discussed and solved all the disagreements together by clarifying the rubrics and identifying more examples of the codes. The six researchers re-graded the same 50 students' responses to those items with unsatisfactory agreements and graded an additional 30 students' responses. The later agreements on 80 students' responses ranged from 74% to 100%. Ten responses still did not exceed 85%. The scorers graded these 10 responses and solved the disagreements through further discussion before re-grading. The recalculated agreements reached 85% or higher. The six researchers then proceeded to grade the remaining students in one of the two assigned booklets.

Data Analysis

In order to examine additional plausible answers to the second research question beyond the preliminary validity evidence in the pilot testing, Rasch models were used for data analysis, with partial credit models applied to the items scored at three levels (i.e. scores 0, 1, and 2). We examined the psychometric properties of the developed items both individually and as a whole. Specifically, the information about item fit was examined to evidence the consistencies between individual items and the models as an indication of the item quality. A multidimensional Rasch model was applied to evaluate the substantive aspect of the unified construct validity, and to examine whether the four components of the inquiry ability construct (i.e. questioning, experimenting, analyzing, and explaining) were distinctive as defined in our framework.

After the examination of dimensionality, we also drew on sources of evidence, which are derived from the Rasch modeling, to evaluate the generalizability aspect of the unified construct validity and to examine the quality of our assessment. Firstly, the person separation reliability was used to check the generalizability across items (Wright & Stone, 1999). Second, a Wright map was constructed to evaluate the alignment between the distributions of student abilities and item difficulties. As suggested by the literature, when the item difficulty is close to the student estimate, the item offers maximum information about the student's ability (de Ayala, 2008). That is, an assessment with too many items at positions far away from the student ability estimates would be too easy or too difficult and therefore would offer little information about students' abilities.

Further evidence of the structural aspect of the unified construct validity was then evaluated by examining the comparability between the scoring rubrics and the framework of inquiry abilities. We calculated mean thresholds for all item responses to investigate the comparability. Another Wright map was constructed to visually present the result of this analysis and evidenced the consistent orders between the levels of performance complexity and the difficulties of those corresponding responses in our assessment.

Results

In order to evaluate the substantive, generalizability, and structural aspects of the unified construct validity of MASIA, Rasch modeling was applied to investigate three issues: (1) the dimensionality of the construct, (2) the internal consistencies of the inquiry items, and (3) the comparability of the scoring rubrics and the inquiry framework. In this section, item fits along with the discriminant indexes are presented to indicate the psychometric properties of the developed items and to confirm the applicability of the Rasch models to the items, individually, followed by an investigation of the first issue of dimensionality. Furthermore, the person separation reliability that was estimated based on the Rasch models is shown to evaluate the internal consistency issue. The alignment between the distributions of student abilities and item difficulties on the Wright map are also presented to indicate the quality of MASIA in terms of measuring the targeted subjects (i.e. students from grades 7 to 12). Finally, a comparison between the scoring rubrics and the inquiry framework is described to explore the third issue of comparability.

Item Fit

Weighted mean squared fit statistics were used to evaluate the fitness of items after they were analyzed by the Rasch models. These statistics are based on the standardized residuals between what is observed and what is expected in the model and were provided by ConQuest (Wu & Adams, 2007). The expected value of this fit statistic would be equal to 1 when the observed responses perfectly match the expected responses. Items with absolute *t*-values greater than 2 are considered an unacceptable fit since this *t*-statistic offers a hypothesis testing framework as to whether the fit statistic deviated from the expected value of 1 (de Ayala, 2008; Wu, Adams, Wilson, & Haldane, 2007). The analysis indicated that slightly more than 72.8% (83) of items had acceptable fit (i.e. t statistics between −2 and 2).

Based on the same criterion, there were, however, 27.2% (31) of items that did not have acceptable fit (two responses were excluded by ConQuest because all students provided exactly the same responses to the two items). A close examination of other psychometric properties was conducted to seek the possible causes of these misfits. Just less than 15.8% (18) of the items had t values of less than −2, possibly implying the high discrimination powers of these items and suggesting steeper slopes of the current students' response curves in comparison to those estimated by the Rasch model (Adams & Khoo, 1996). Figure 4(a) shows a cumulative probability curve of such an item with category one scored as one and category two scored as two. The misfit resulted from a greater increase (or decrease) in observed probability when the theoretical probability increased (or decreased). A discrimination index below .25 is an advised cutoff criterion to signal those items with insufficient powers to discriminate among students (Adams & Wu, 2002). We conducted follow-up examinations of these items. No items had a discrimination index smaller than the criterion, which therefore, confirmed this speculation. The items with high

discriminant powers were retained since they could serve as tools to distinguish between high- and low-performing students.

Slightly more than 12.3% of the items (13 in total), however, had *t*-values greater than 2, suggesting misfits in these responses. Among these 13 responses, 9 items also showed poor discrimination powers (less than .25). Figure 4(b) presents a

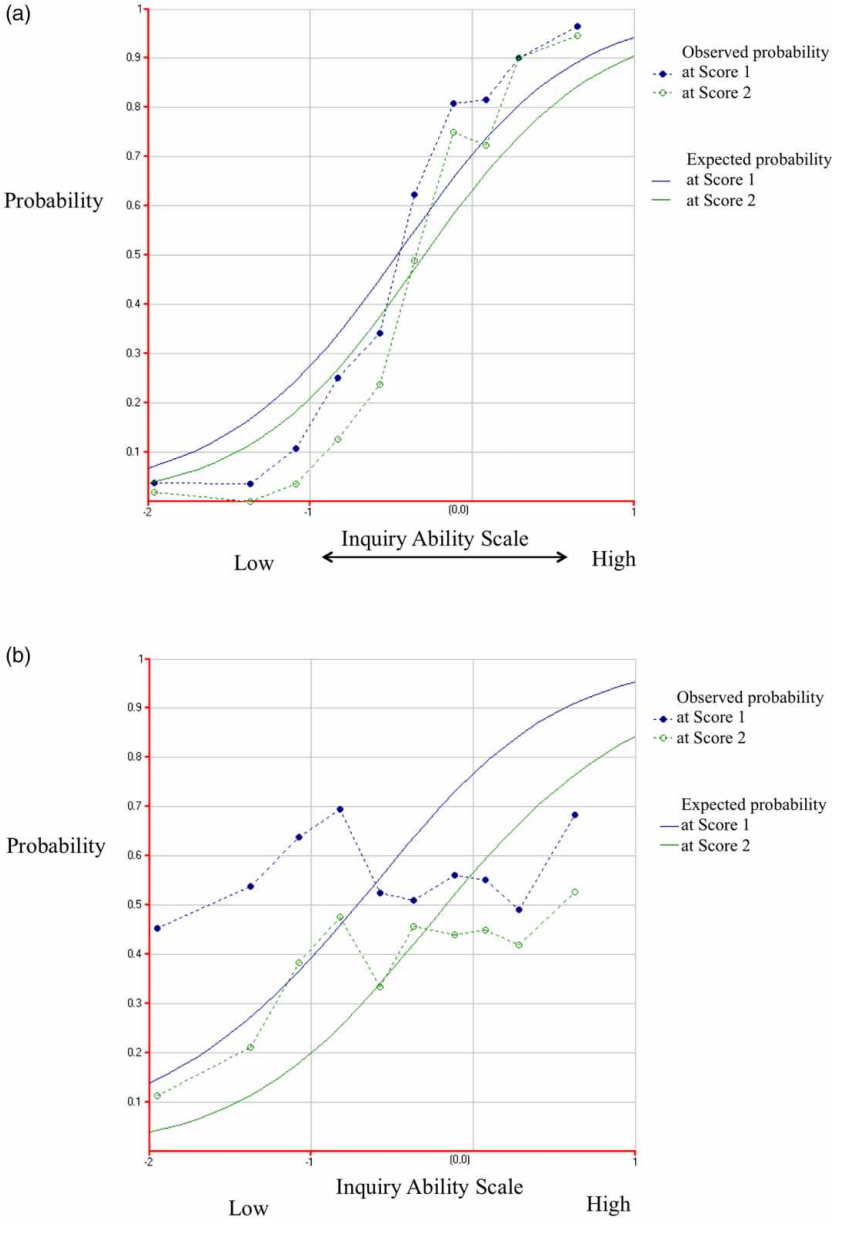

Figure 4. The cumulative curves with observed probabilities and expected probabilities from the Rasch model for a misfit item (a) with high discrimination and (b) with low discrimination

cumulative probability curve of such an item. The misfit resulted from a flat slope in the students' response curve. That is, the observed probabilities were not higher for students with higher estimated abilities, and were not lower for students with lower abilities. The misfit in the figure might also signify a reversed pattern between the overall estimated abilities and the earned scores for the items. This reversed pattern was evident in the frequency table when the scored responses were linked back to the students' abilities (see Table 4). As indicated in Table 4, students whose responses with a higher score (e.g. 1) had a lower average of estimated abilities than those of the group with a lower score (e.g. 0). An examination of all 13 responses revealed that 6 of these 9 items also entailed reversed patterns. Regardless of the patterns, the nine items of these responses were all excluded from the item pool to ensure the quality of the items in our assessment.

Additionally, among the 13 responses with misfits, a 5-question task was removed from the item pool. Three considerations guided this decision. Firstly, one item in the same task was excluded because of the poor psychometric properties discussed above. Secondly, two of the five items revealed reversed patterns. Thirdly, all five items were too difficult for students in this testing, with only 0.6% of the students (6 out of 1,066) who had estimated abilities higher than the easiest item (based on calibrated item difficulty) among these five. These items therefore did not serve as a good tool to reveal information about students' inquiry abilities.

In sum, the results of the item analysis indicated that most of the items entailed satisfactory psychometric properties and were able to discriminate students with higher levels of proficiency from those with lower levels.

Dimensionality of the Construct

In order to investigate the substantive aspect of the unified construct validity, the latent structure of the construct of scientific inquiry abilities in our assessment was examined by fitting with a four-dimensional model. The four-dimensional model informed how well our assessment would measure the four inquiry abilities as four distinctive components. The results showed that the person separation reliabilities were .85, .88, .83, and .87 for the questioning, experimenting, analyzing, and explaining abilities, respectively. These reliabilities are conventionally interpreted as the adequacy to discriminate students by the measures of each of the four inquiry abilities (Wright & Stone, 1999).

Table 4. Response frequency and averaged ability estimate in each scored point in an item with misfit

Score	Frequency	Percentage	Averaged ability estimate
0	230	42.98	−0.60
1	96	18.36	−0.86
2	197	37.67	−0.37

However, a close examination of the four dimensions revealed that the correlation coefficients between the four dimensions ranged from .87 to .96 (Table 5). The contrast between the correlations and the person separation reliabilities suggested imprecise estimates of four abilities because of larger measurement errors (indicated by the reliabilities), indicating distinctions among the measures of these four abilities. Therefore, the results implied that our assessment offered more precise estimates when the assessment was intended to obtain information about students' inquiry abilities as one coherent construct.

Alignment between Student Ability and Item Difficulty

After examination of the psychometric properties and dimensionality of the developed items, 101 items in 29 tasks were kept in the item pool. Another item calibration based on the same unidimensional Rasch model was conducted, and the re-modeled analysis indicated that the person separation reliability was .91, which added evidence to support the generalizability aspect of the unified construct validity. A Wright map was created based on the 101 items in the 29 tasks and depicted both student ability distribution and item difficulty distribution on the same metric for direct comparison. Figure 5 illustrates the Wright map of our assessment. Each *x* in Figure 5 represents 6.5 students, and its position indicates that these students had a .5 probability of offering correct responses to the items in the same position. The Wright map shows that the items in our assessment distributed similarly to the targeted students and covered the range of the student estimates. Therefore, the tool could provide accurate measures for 8th and 11th graders. That is, our assessment could offer accurate information about most high school students' inquiry abilities, providing evidence of the quality of our assessment.

Yet, relatively few items were below −2.25 in contrast to the number of students in this range. On the contrary, there were only a few students relative to items with difficulty estimates above 1.38. The patterns thus uncovered two issues: (1) relatively few items provided more precise information about students at lower levels, particularly those below −2.25; (2) there were too many difficult items at the higher end of the continuum, particularly for those with difficulty above 1.38. In the next version of MASIA, we will address the former issue by developing additional items to achieve a balanced design of item difficulty. Particularly, items calling for selected responses and

Table 5. Estimated correlations among the four latent constructs in a four-dimensional rasch model

Inquiry abilities	Questioning	Experimenting	Analyzing	Explaining
Questioning	–			
Experimenting	.94	–		
Analyzing	.87	.91	–	
Explaining	.94	.96	.90	–

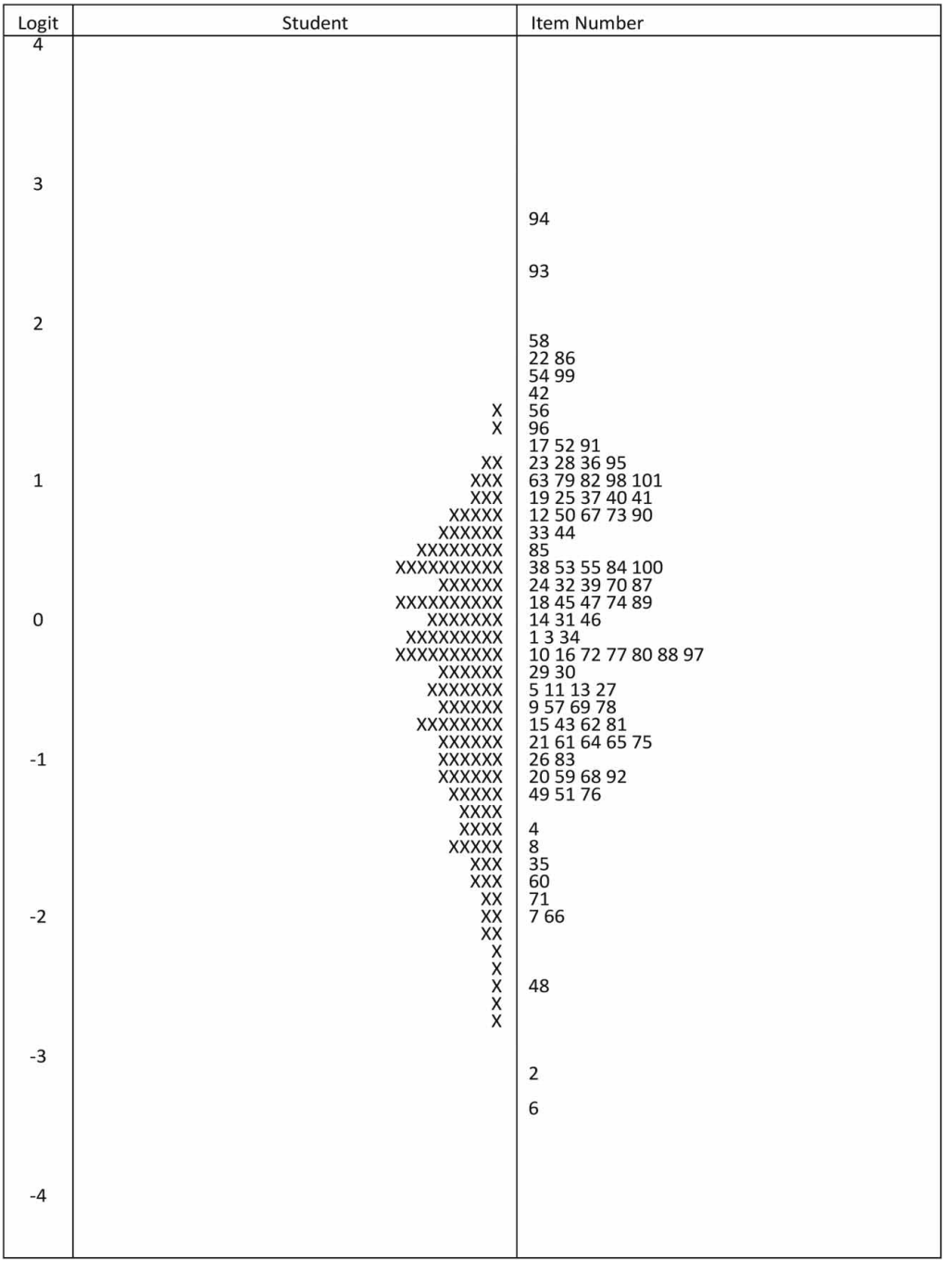

Figure 5. Wright map of the 101 MASIA items. Each *x* represents 6.5 students

requiring less sophisticated performances will be included as informed by the three items with difficulty below −2.25.

However, despite the latter issue, we did not remove the items with difficulty estimates higher than 1.38 since it is possible that more students could achieve the higher score levels when inquiry abilities are more highly valued in schools in Taiwan or when MASIA is administered to more adept students. The same assessment can be applied to this future condition.

Comparability between Scoring Rubrics and the Framework of Inquiry Abilities

To evaluate the structural aspect of the unified construct validity, we examined the extent to which our scoring rules and grading rubrics were comparable with the framework of inquiry abilities. Mean thresholds of item responses at three performance levels across the inquiry sub-abilities were calculated. Table 6 shows that the mean thresholds increased as the responses reflected higher performance levels in item responses from the 12 sub-abilities. This suggests that our items offered satisfactory information about different levels of student performance complexity. The comparable orders between the assigned scores and the levels of performance complexity thus supported the alignment between our scoring rubrics and our inquiry framework.

However, among the 12 sub-abilities in Figure 6, the mean thresholds of the *transforming data* sub-ability of Analyzing (A2) between middle and high levels of

Table 6. Mean thresholds of item responses at three performance levels across the inquiry sub-abilities

	Level of performance complexity					
	Low		Middle		High	
Inquiry sub-abilities	*M*	*n*	*M*	*n*	*M*	*n*
Questioning						
Q1			−0.53	3	0.76	3
Q2			−2.40	1	−0.20	2
Q3			−0.33	15	1.90	7
Experimenting						
R1			0.16	18	1.82	2
R2			−0.64	8	0.87	9
R3			−0.20	1	1.34	5
Analyzing						
A1			−1.29	11		
A2	−0.43	6	0.24	8	0.35	2
Explaining						
E1			−1.34	12	0.37	10
E2			−1.07	3	0.47	6
E3			−0.13	8	1.28	8
E4	−0.46	4	0.94	14	4.64	2

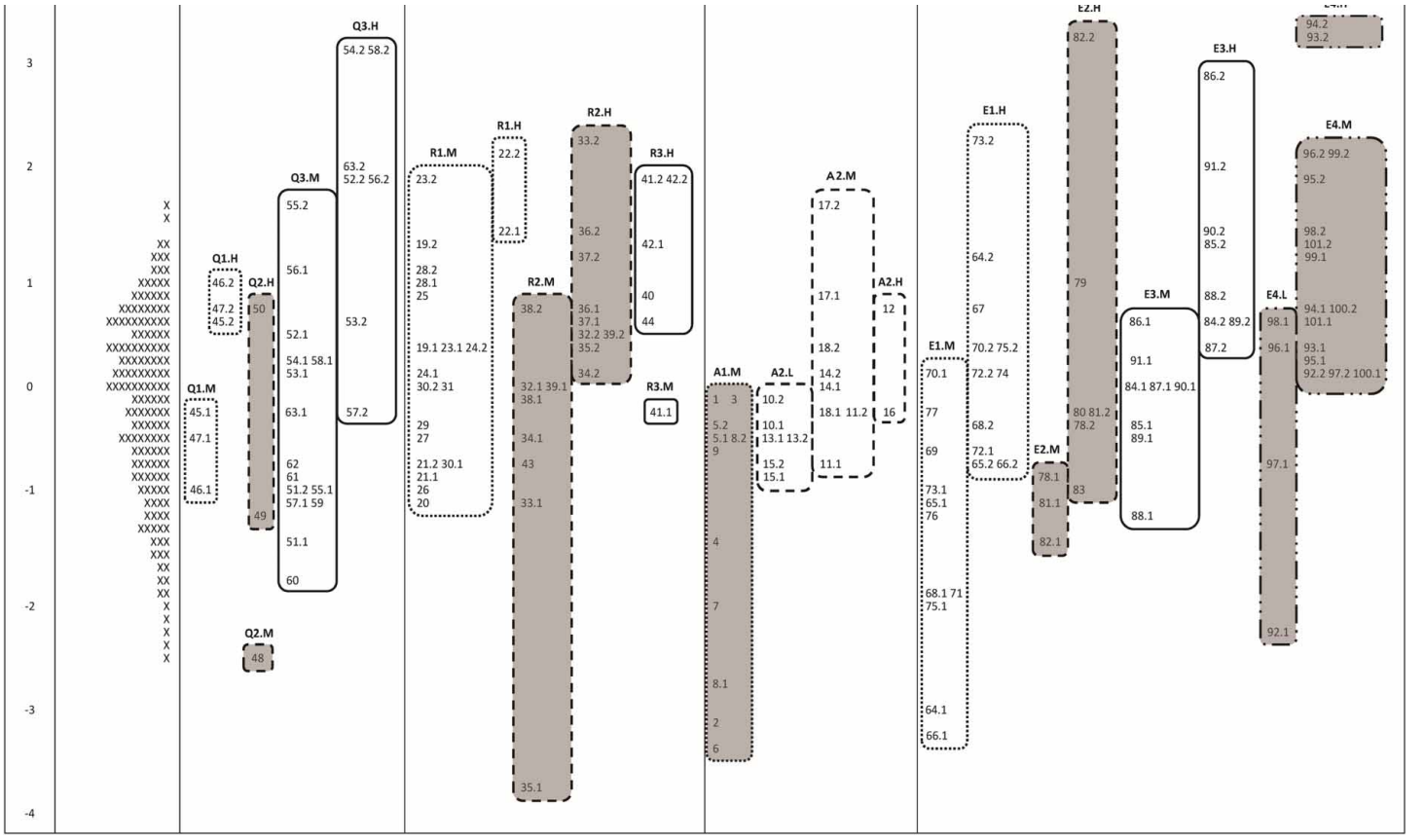

Figure 6. Wright map for items of inquiry sub-abilities in MASIA, showing the student distribution and the item response thresholds. Each *x* represents 6.5 students and each number represents the item number followed by a dot and a digit to indicate the responses earning full or partial credit scores if they are available. Q1 = formulating research questions; Q2 = identifying research questions; Q3 = making predictions; R1 = identifying controlled and manipulated variables; R2 = planning experimental procedures; R3 = selecting appropriate measurements; A1 = identifying relevant data; A2 = transforming data; E1 = making a claim; E2 = using evidence; E3 = reasoning from evidence to the claim; E4 = offering and evaluating alternative explanations. L = low level of performance complexity; M = middle level of performance complexity; H = high level of performance complexity

performance complexity (A2.M and A2.H) did not indicate a discernible difference, although mean thresholds at the high level of performance complexity ($M = 0.35$) were higher than those at the middle level ($M = 0.24$) in Table 5. That is, the difficulties of two item responses from the most difficult item (item 17; 17.2 and 17.1) of the middle level were even higher than the two items measuring the high level of the sub-ability (items 12 and 16). A close examination of these items indicated that both responses in item 17 required a graphical transformation for three variable relationships, while items 12 and 16 asked for mathematical transformations for bivariate relationships. This suggests that, in addition to the types of transformation between data and representations, the number of variables involved in the data also seemed to be a factor affecting the item difficulty.

Additionally, item responses of eight sub-abilities (i.e. Q3, R1, R2, A2, E1, E2, E3, and E4) presented overlapping patterns of difficulty thresholds between the successive levels of performance (Figure 6). That is, some item responses at the higher performance levels were easier than some at the lower performance levels even though the mean item difficulties increased as the performance levels increased. Together the item analysis and the overlapping pattern suggested that MASIA contains items with various characteristics, and thus allows students to apply their inquiry abilities in different situations. The items in MASIA thereby enabled detection of more fine-grained differences in the inquiry proficiency levels.

Discussion

The purpose of this study was to develop and validate a multimedia-based assessment of scientific inquiry abilities that covers a more comprehensive construct of inquiry abilities and which targets secondary school students in different grades while authentically measuring their inquiry abilities. By applying the construct modeling approach (Brown & Wilson, 2011; Shin et al., 2010; Wilson, 2005), we employed five design steps to guide the assessment design process, and developed a valid assessment of secondary students' inquiry abilities. Below, we discuss the issues in designing a valid multimedia-based assessment of scientific inquiry abilities with the potential for offering comprehensive and continuous information. We also indicate the limitations and future applications of the study.

Issues of Scientific Inquiry Abilities as a Unidimensional Construct

Previous research has shown that when students demonstrated successfully inquiry performances, there were still variations in their profiles across student groups, even for the same tasks (Kind, Kind, Hofstein, & Wilson, 2011; Wu & Hsieh, 2006). This implies the distinctiveness of students' scientific inquiry abilities; different inquiry abilities could be viewed as having multiple dimensions. However, the possibility of these abilities being a unidimensional construct in an inquiry assessment has not been ruled out (Kind, 2013). In our results, the correlations among the four inquiry components were higher than the person separation reliabilities of the four

components, which signaled imprecise estimates from the four components. This study suggested the more parsimonious unidimensional structure as one candidate for our assessment; the results were not in accordance with a multidimensional inquiry construct. Future studies may design an assessment with more items to explore whether the inquiry abilities are distinctive as four constructs.

Issues of Item Characteristics and Item Difficulty

Our results also suggested that other item characteristics likely accounted for the item difficulty above and beyond the power of the sub-abilities in our framework. Specifically, response formats and the numbers of variables involved in a task appeared to affect the item difficulty in this study.

In our assessment, items with difficulty below −2.25 all called for selected responses but not constructed responses. Selected responses asking for identification from alternatives likely demanded less than constructed responses calling for a product or demonstration of inquiry practices when the same items employed these two types of response formats (Lee & Liu, 2009). By employing selected responses, our assessment could tap rudimentary understanding of inquiry practices for those who were starting to learn and develop scientific inquiry abilities, and were sensitive to the differences between those students with low levels of ability.

The number of variables involved in a task also seemed to be associated with item difficulty in multiple inquiry abilities within our framework. In our assessment, of two of the three items with difficulties below −2.25, one called for identifying appropriate data in the item of Analyzing and the other asked for selecting feasible questions in the item of Questioning, and all the alternative options in these two items involved bivariate rather than multivariate relationships. By contrast, items requesting graphical transformation for three-variable relationships were more difficult than those requiring the same transformation for bivariate relationships. Possibly, successful scientific inquiry practices involving multivariate relationships demand integrated understandings of multivariable causality that are not required for bivariate reasoning (Kuhn, 2007; Wu, Wu, Zhang, & Hsu, 2013). Future research may clarify how these two characteristics may affect item difficulty and whether the effects can be generalized to all inquiry abilities. By doing so, implications for how to design items of appropriate difficulty levels can be made, particularly for those assessments that are intended to serve a wide range of students and offer information about learning progression in inquiry across all secondary years.

Limitations of the Study

Although this study extends the application of CBAs to offering summative, accountability information for students of different grades beyond the recent research efforts, there are some limitations. Firstly, students' responses on MASIA were mainly from the products of inquiry activities with the simulations. Yet, the processes of manipulating the simulations as additional evidence of students' proficiency may increase the

fidelity to our inquiry ability construct. A few efforts have been made to establish valid scoring of such complex process outcomes (e.g. Gobert et al., 2013; Zoanetti, 2010) but are not in widespread use, possibly due to the issue of validity when the automated scoring is applied. Future research may explore the possibility of using the processes as evidence of students' proficiencies in science, and facilitate the further application of CBAs.

The second limitation is pertinent to students' science content knowledge in their inquiry performance. In MASIA, the tasks that were designed around various science topics in four content areas could increase the fidelity to the integrated view of inquiry abilities and allow students to 'mesh the [science] processes with scientific knowledge' (NRC, 2000, p. 18) across four science content areas. Yet, the measures from such tasks may reflect levels of students' scientific inquiry abilities as well as their content knowledge. Future studies may collect information of students' content knowledge not only in these four content areas but also in those various topics to understand the relationships between the inquiry measures in MASIA (or other CBAs) and science content knowledge. By doing so, future studies can explore the ways to fully control the confounding effects of content knowledge while keeping the fidelity to the integrated view of inquiry abilities in the assessment.

Finally, the items in MASIA were organized around testlets. That is, a group of items were related to a common content area and shared the same task situations, topics and item stems. Student inquiry performances were not independent and were possibly influenced by their performance in the same testlets. Because of this organization, we cannot be certain how pervasive the testlets' effect on the students' performance was. Nor can we be certain whether the effects vary by the students' characteristics such as their test taking styles, possibly contributing to a potential threat to MASIA. Yet, the organization around testlets saved testing time and alleviated the students' cognitive load because they only had to comprehend a limited number of contexts. Furthermore, engaging in a series of inquiry activities is more authentic than performing discrete and isolated tasks that demand only one inquiry ability or even one sub-ability at a time. Future studies may examine whether the testlet organization would interact with other factors to confound the inquiry measures.

Future Applications of the Study

The development and validation in this study bodes well for further applications of MASIA. Given the inadequacy of authentic measures of inquiry in large-scale accountability assessments (Quellmalz & Pellegrino, 2009), MASIA can serve as a credible tool for accountability purposes to inform the quality of inquiry teaching and learning in schools. Additionally, schools and classroom teachers can select, link, and sequence level- and grade-appropriate inquiry tasks and items in MASIA from different science topics and content areas along with the scaled item difficulties for their internal assessment systems. The coordinated inquiry tasks in the systems will assist teachers in identifying students' mastery levels in inquiry to pace their instructional sequences

across curriculum units and areas. Furthermore, individual teachers can integrate the inquiry tasks and items of MASIA into their teaching and assessment practices for those topics covered in MASIA. The selected tasks will allow students to learn the topics in an authentic way and help teachers offer just-in-time interventions.

To further facilitate student inquiry learning in classrooms, assessment designers can advance applications of MASIA for individualized and/or curriculum-embedded testing (Kuo & Wu, 2013). Assessment designers can enhance the capacities of MASIA to gather both the processes and products of students' inquiry activities, and to enable automatic evaluation. The automated scoring of the processes and products will allow MASIA to capture rich and complex data of students' inquiry, thus providing insights into student learning without interrupting their assessment activities. With unobtrusive and continuous evaluation, the inquiry tasks can be extended to individualized tasks that provide immediate feedback/hints to meet each student's needs, and to curriculum-embedded assessments that offer guidance to scaffold students' inquiry learning.

Conclusion

In conclusion, our five design steps ensured that MASIA could offer valid information about students' scientific inquiry abilities in a comprehensive and continuous way. The five design steps began with the construct-driven framework from relevant theories and enabled MASIA's coherence by aligning the item and task design, the scoring rubric construction, the pilot testing, and the Rasch modeling with the framework. The five design steps thus suggest a theory-driven procedure for assessment designers who intend to materialize the potentials of CBAs.

During the implementation, multiple sources of evidence in the pilot testing and Rasch modeling were collected to support the unified construct validity of MASIA. In accordance with Messick's unified conception of construct validity, the empirical evidence suggests a comprehensive and integrative way for assessment designers to address a CBA's validity from multiple but interrelated aspects.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

This study was based upon work supported by the Ministry of Science and Technology in Taiwan under MOST 103-2511-S-003-038-MY4, MOST 103-2811-S-003-010, and the Aim for the Top University Project at the National Taiwan Normal University.

ORCID

Hsin-Kai Wu http://orcid.org/0000-0003-0018-9969

References

Abd-El-Khalick, F., Boujaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., … Tuan, H. L. (2004). Inquiry in science education: International perspectives. *Science Education*, *88*, 394–419.

Adams, R. J., & Khoo, S. T. (1996). *Quest: The interactive test analysis system*. Camberwell, Australia: Australian Council for Educational Research.

Adams, R. J., & Wu, M. L. (2002). *PISA 2000 technical report*. Paris: OECD.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standard for educational psychological testing*. Washington, DC: American Educational Research Association.

de Ayala, R. J. (2008). *Methodology in the social sciences: Theory and practice of item response theory*. New York, NY: Guilford Press.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automad scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, *17*(4), 9–17. doi:10.1111/j.1745-3992.1998.tb00631.x

Bennett, R. E., Persky, H., Weiss, A., & Jenkins, F. (2010). Measuring problem solving with technology: A demonstration study for NAEP. *The Journal of Technology, Learning and Assessment*, *8* (8). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1627/1471

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*(1), 33–63. doi:10.1207/s15326977ea1101_2

Brown, N. J. S., & Wilson, M. (2011). A model of cognition: The missing cornerstone of assessment. *Educational Psychology Review*, *23*(2), 221–234. doi:10.1007/s10648-011-9161-z

Buckley, B. C., Gobert, J. D., Horwitz, P., & O'Dwyer, L. M. (2010). Looking inside the black box: assessing model-based learning and inquiry in BioLogica™. *International Journal of Learning Technology*, *5*(2), 166–190.

Chung, G., de Vries, L. F., Cheak, A. M., Stevens, R. H., & Bewley, W. L. (2002). Cognitive process validation of an online problem solving assessment. *Computers in Human Behavior*, *18*(6), 669–684.

Crundwell, R. M. (2005). Alternative strategies for large scale student assessment in Canada: Is value-added assessment one possible answer. *Canadian Journal of Educational Adminstration and Policy*, *41*, 1–21.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academy Press.

Frederiksen, J. R., & White, B. Y. (1998). Teaching and learning generic modeling and reasoning skills. *Interactive Learning Environments*, *5*(1), 33–51.

Garden, R. A. (1999). Development of TIMSS performance assessment tasks. *Studies in Educational Evaluation*, *25*(3), 217–241. doi:10.1016/S0191-491X(99)00023-1

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, *22*(4), 521–563. doi:10.1080/10508406.2013.837391

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IEA-ETS Research Institute Monograph*, *3*, 125–156.

Kind, P. M. (2013). Establishing assessment scales using a novel disciplinary rationale for scientific reasoning. *Journal of Research in Science Teaching*, *50*(5), 530–560.

Kind, P. M., Kind, V., Hofstein, A., & Wilson, J. (2011). Peer argumentation in the school science laboratory: Exploring effects of task features. *International Journal of Science Education*, *33*(18), 2527–2558. doi:10.1080/09500693.2010.550952

Krajcik, J. S., Blumenfeld, P. C., Marx, R. W., Bass, K. M., Fredricks, J., & Soloway, E. (1998). Inquiry in project-based science classrooms: Initial attempts by middle school students. *Journal of the Learning Sciences*, *7*(3&4), 313–350.

Krajcik, J. S., & Czerniak, C. M. (2007). *Teaching children science in elementary and middle school: A project-based approach*. New York, NY: Routledge.

Kuhn, D. (2007). Reasoning about multiple variables: Control of variables is not the only challenge. *Science Education*, *91*, 710–726.

Kuo, C.-Y., & Wu, H.-K. (2013). Toward an integrated model for designing assessment systems: An analysis of the current status of computer-based assessments in science. *Computers & Education*, *68*, 388–403. doi:10.1016/j.compedu.2013.06.002

Lee, H.-S., & Liu, O. L. (2009). Assessing learning progression of energy concepts across middle school grades: The knowledge integration perspective. *Science Education*, *94*(4), 665–688. doi:10.1002/sce.20382

Lee, H.-S., Liu, O. L. & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education*, *24*(2), 115–136.

Liu, O. L., Lee, H., Hofstetter, C. & Linn, M. C. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. *Educational Assessment*, *13*, 33–55.

Lorch, R. F., Lorch, E. P., Calderhead, W. J., Dunlap, E. E., Hodell, E. C., & Freer, B. D. (2010). Learning the control of variables strategy in higher and lower achieving classrooms: Contributions of explicit instruction and experimentation. *Journal of Educational Psychology*, *102*(1), 90–101. doi:10.1037/a0017972

Martin, M. O., & Mullis, I. V. S. (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Ministry of Education. (1999). *Curriculum outlines for "nature science and living technology"*. Taipei, Taiwan: Ministry of Education.

Ministry of Education. (2008). *Curriculum outlines for senior high schools*. Taipei, Taiwan: Ministry of Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50* (9), 741–749. doi:10.1037/0003-066X.50.9.741

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

National Research Council. (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. Washington, DC: National Academy Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. In J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.), Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academies Press.

Organization for Economic Co-operation and Development. (2010). *PISA computer-based assessment of student skills in science*. Paris: Author.

Pine, J., Aschbacher, P., Roth, E., Jones, M., McPhee, C., Martin, C., … Foley, B. (2006). Fifth graders' science inquiry abilities: A comparative study of students in hands-on and textbook curricula. *Journal of Research in Science Teaching*, *43*(5), 467–484. doi:10.1002/tea.20140

Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and testing. *Science*, *323*(5910), 75–79.

Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., & Buckley, B. C. (2012). Science assessments for all: Integrating science simulations into balanced state science assessment systems. *Journal of Research in Science Teaching*, *49*(3), 363–393. doi:10.1002/tea.21005

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching*, *33*(10), 1045–1063.

Shin, N., Stevens, S. Y., & Krajcik, J. (2010). Tracking student learning over time using construct-centered design. In S. Rodrigues (Ed.), *Using analytical frameworks for classroom research: Collecting data and analysing narrative* (pp. 38–68). London: Taylor & Francis.

Songer, N. B., Lee, H. S., & McDonald, S. (2003). Research towards an expanded understanding of inquiry science beyond one idealized standard. *Science Education*, *87*(4), 490–516.

Wenning, C. J. (2007). Assessing inquiry skills as a component of scientific literacy. *Journal of Physics Teacher Education Online*, *4*(2), 21–24.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.
Wright, B. D. & Stone, M. H. (1999). *Measurement essentials*. Wilmington, DE: Wide Range.
Wu, M. L., & Adams, R. J. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement Solutions.
Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software*. Melbourne: Australian Council for Educational Research.
Wu, H.-K., & Hsieh, C.-E. (2006). Developing sixth graders' inquiry skills to construct scientific explanations in inquiry-based learning environments. *International Journal of Science Education*, *28*(11), 1289–1313. doi:10.1080/09500690600621035
Wu, P. H., Wu, H.-K., & Hsu, Y. S. (2014). Establishing the criterion-related, construct, and content validities of a simulation-based assessment of inquiry abilities. *International Journal of Science Education*, *36*(9–10), 1630–1650. doi:10.1080/09500693.2013.871660.
Wu, H.-K., Wu, P. H., Zhang, W. X., & Hsu, Y. S. (2013). Investigating college and graduate students' multivariable reasoning in computational modeling. *Science Education*, *97*, 337–366. doi:10.1002/sce.21056.
Zachos, P., Hick, T. L., Doane, W. E., & Sargent, C. (2000). Setting theoretical and empirical foundations for assessing scientific inquiry and discovery in educational programs. *Journal of Research in Science Teaching*, *37*(9), 938–962. doi:10.1002/1098-2736(200011)37:9<938::AID-TEA5>3.0.CO;2-S
Zoanetti, N. (2010). Interactive computer based assessment tasks: How problem-solving process data can inform instruction. *Australasian Journal of Educational Technology*, *26*(5), 585–606.