Routledge
Taylor & Francis Group

International Journal of Science Education

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

Development and validation of an instrument for evaluating inquiry-based tasks in science textbooks

Wenyuan Yang & Enshan Liu

To cite this article: Wenyuan Yang & Enshan Liu (2016): Development and validation of an instrument for evaluating inquiry-based tasks in science textbooks, International Journal of Science Education, DOI: 10.1080/09500693.2016.1258499

To link to this article: http://dx.doi.org/10.1080/09500693.2016.1258499

Published online: 20 Dec 2016.

INTERNATIONAL JOURNAL OF SCIENCE EDUCATION, 2016
http://dx.doi.org/10.1080/09500693.2016.1258499

Routledge
Taylor & Francis Group

Development and validation of an instrument for evaluating inquiry-based tasks in science textbooks

Wenyuan Yang[a] and Enshan Liu[b]

[a]School of Life Sciences, Capital Normal University, Beijing, People's Republic of China; [b]College of Life Sciences, Beijing Normal University, Beijing, People's Republic of China

ABSTRACT

This article describes the development and validation of an instrument that can be used for content analysis of inquiry-based tasks. According to the theories of educational evaluation and qualities of inquiry, four essential functions that inquiry-based tasks should serve are defined: (1) assisting in the construction of understandings about scientific concepts, (2) providing students opportunities to use inquiry process skills, (3) being conducive to establishing understandings about scientific inquiry, and (4) giving students opportunities to develop higher order thinking skills. An instrument – the Inquiry-Based Tasks Analysis Inventory (ITAI) – was developed to judge whether inquiry-based tasks perform these functions well. To test the reliability and validity of the ITAI, 4 faculty members were invited to use the ITAI to collect data from 53 inquiry-based tasks in the 3 most widely adopted senior secondary biology textbooks in Mainland China. The results indicate that (1) the inter-rater reliability reached 87.7%, (2) the grading criteria have high discriminant validity, (3) the items possess high convergent validity, and (4) the Cronbach's alpha reliability coefficient reached 0.792. The study concludes that the ITAI is valid and reliable. Because of its solid foundations in theoretical and empirical argumentation, the ITAI is trustworthy.

ARTICLE HISTORY
Received 17 April 2016
Accepted 4 November 2016

KEYWORDS
Inquiry-based task; science textbook; content analysis; scientific inquiry

Introduction

The development of textbooks that foster inquiry is an emphasis of curricular reform in K-12 science because inquiry has become a major focus of science education over the past few decades (Meyer, Meyer, Nabb, Connell, & Avery, 2013; Shulman & Tamir, 1973; Trumbull, Bonney, & Grudens-Schuck, 2005). Science is not only a body of knowledge that mirrors what we know about the world but also a set of multiple manners in which scientists study the world (National Research Council [NRC], 1996, 2000, 2012). Scientific inquiry refers to the approach scientists use and reflects how science proceeds. A lack of experience with scientific inquiry precludes deep understandings about scientific ideas. Providing authentic opportunities for students to engage in scientific inquiry is conducive to enhancing their abilities to develop rational explorations and explanations of their world. Consequently, scientific inquiry is a highly recommended learning process

CONTACT Enshan Liu ✉ liues@bnu.edu.cn College of Life Sciences, Beijing Normal University, No. 19 Xinjiekouwai Street, Haidian District, Beijing 100875, People's Republic of China

in which students propose ideas based on evidence derived from their practices, develop understandings about scientific concepts, and make sense of how to engage in science (Anderson, 2002; Bybee, 2010; Schwab, 1962).

Since the phrase 'teaching science as enquiry' was first developed by Schwab in 1962 at Harvard University, many changes and innovations have been brought to K-12 science textbooks (Chiappetta, 2008). A diversity of inquiry-based tasks has appeared in science textbooks to guide students in conducting scientific inquiry, such as experiments, laboratory activities, investigations, and practical work. Today in most K-12 classrooms, textbooks serve as the principal tool and tutor of teaching and learning and have an enormous influence on what is taught in science classrooms and how the curriculum is presented (McDonald, 2016; Pingel, 2010; Roseman, Kulm, & Shuttleworth, 2001). It is assumed that well-designed inquiry-based tasks in science textbooks play an important role in supporting students' experience with scientific inquiry and developing understandings about scientific ideas. In turn, the evaluation of inquiry-based tasks in science textbooks deserves serious concern.

Criteria for an evaluation of the design quality of inquiry-based tasks

Since Ralph W. Tyler proposed the concept of educational evaluation, such evaluation has always been purpose-driven (Cronbach et al., 1980; Sax & Newton, 1997; Tyler, 1942). Generally speaking, evaluation is making value judgments. In the field of education, evaluation refers to conducting systematic investigations involving goals, procedures, principles, and tools to measure the value or quality of education-related matters such as curriculum, instructional approaches, learning materials, educators, and learners. Accordingly, content analysis of textbooks is essentially the measurement of textbooks' value. The textbook is the principal tool of teaching and learning in K-12 science classrooms, and the core value of a textbook is embodied in how well the book functions as an instructional tool. Therefore, content analysis of inquiry-based tasks in science textbooks is a measurement of the degree to which these tasks perform their functions.

The first requirement is a definition of functions that inquiry-based tasks in science textbook should serve. With regard to the functions of inquiry-based tasks, it is better to search explanations of scientific inquiry for the answer. Scientific inquiry is also known as laboratory activities. Shulman and Tamir (1973) emphasised that the laboratory has always been the most distinctive feature of science instruction and is central to the science learning process. Those authors listed five groups of objectives that students may achieve in laboratory-based science classes: skills, concepts, cognitive abilities, understanding the nature of science, and attitudes. Bybee (2002, 2006) interpreted scientific inquiry to be a set of particular methods with which scientists explore the natural world, such as observations and experiments that result in empirical evidence used to answer a scientific question. In addition, scientific inquiry is not only the science process that students are encouraged to experience but also the content knowledge of science curriculum that students are expected to understand. An inquiry-based learning process enables students to develop fundamental abilities and construct conceptual understandings about scientific inquiry. Lederman and Lederman (2012) suggested that scientific inquiry includes traditional science processes such as questioning, predicting, observing, analysing data, inferring, and interpreting but also refers to the combining of these processes with scientific knowledge, scientific reasoning, and critical thinking to develop scientific knowledge. In the report of inquiry-based science education (IBSE) programmes published by the InterAcademies Panel (2006), scientific inquiry is defined as students developing skills of inquiry and understandings about scientific concepts by their own activities, which involve direct exploration and experimentation, argumentation, and critical and logical reasoning regarding evidence the students have gathered. It is commonly believed that K-12 classrooms are designed to present established understandings, not to promote the discovery of new knowledge; a school classroom is not a research laboratory. Thus, teaching and learning science as inquiry refers to capturing typical characteristics of professional science within the K-12 school classroom such as problem-solving approaches, scientific reasoning and arguments, and critical thinking skills (Hanauer et al., 2006). In other words, the goals of inquiry in K-12 classrooms are not to study the unknown world, but to construct understandings of what students are expected to learn. Thus, scientific inquiry is at the heart of science learning, engaging students in the process of scientific research. Previous descriptions of scientific inquiry suggest four essential functions that inquiry-based tasks in science textbooks should serve:

(1) to assist in the construction of understandings about scientific concepts;
(2) to provide students opportunities to use inquiry process skills;
(3) to contribute to the establishment of understandings about scientific inquiry; and
(4) to provide students with opportunities to develop higher order thinking skills (HOTS).

In terms of an evaluation of the quality of inquiry-based tasks, the criterion is the degree to which these tasks perform the above functions. In addition, for the purpose of measuring the design quality of inquiry-based tasks, an instrument developed for content analysis should be valid for gathering full and accurate evidence to assess the function performance of a task.

Emergence of inquiry-based textbooks and studies on assessing those textbooks

The explosion of extensive focus on assessing inquiry-based tasks in science textbooks can be traced to the mid-1970s. In the ten years following the curriculum reform of K-12 science education in the mid-1960s, numerous inquiry-based textbooks were rapidly developed and widely adopted; however, independent evaluative efforts lagged behind (Herron, 1971). Even the authors of those textbooks did not have an explicit vision of students' performance expectations after performing inquiry-based tasks. Two factors contributed to this gap: the ambiguous understandings about scientific inquiry and the lack of explicit criteria for evaluating the inquiry in textbooks. Since that time, three types of research addressing the evaluation of inquiry in textbooks have developed.

Studies on assessing the instruction of inquiry process

The first category of studies on evaluating inquiry-based tasks in science textbooks concentrates on the instruction of the inquiry process. Tamir and Lunetta (1978, 1981) developed the Laboratory Structure and Task Analysis Inventory (LAI) on the basis of Schwab and Herron's conceptual framework of inquiry. They harbour the idea that content analysis is a powerful approach that enables the researcher to ascertain the quality of a textbook. The LAI is an instrument for the content analysis of inquiry-based tasks and contains two sections: *Laboratory Organization*, comprising 14 items in four sub-sections, and *Laboratory Tasks*, comprising 24 items in four sub-sections. The LAI has been used to review inquiry-based tasks in several series of biology, physics, and chemistry textbooks. Germann, Haskins, and Auls (1996) modified Tamir and Lunetta's LAI and proposed the Biology Laboratory Manual Inventory (BLI). They used the BLI to evaluate inquiry-based tasks in nine senior secondary biology textbooks. The BLI contains five sections, *Prelab Activity Provided*, *Student Planning and Design*, *Student Performance*, *Student Analysis/ Interpretation*, and *Student Application*. Compared with the LAI, the BLI focused more on the disposition of using scientific processing skills. This category of studies examines the design of laboratory processes to determine how well those processes promote the process skills that are involved in scientific inquiry. Although science is not simply a body of knowledge, neither is it merely a suite of practices. Coupling practice with knowledge provides the learning context, whereas practices alone are activities and knowledge alone is memorisation. Both the LAI and the BLI barely mention the understandings about scientific knowledge, which separates the scientific process from scientific content. In addition, instructing students to perform an inquiry process is one of the four functions of inquiry-based tasks whereas the other three functions are missing. Because these instruments focus only on the inquiry process, this category of studies is unable to evaluate an inquiry-based task comprehensively.

Studies on discriminating the openness levels of inquiry

The second category of studies on evaluating inquiry-based tasks in science textbooks seeks to distinguish the openness levels of inquiry, which is also the most common mode of analysing inquiry-based tasks. Schwab (1962) suggested that inquiry can be divided into three phases: problems, methods, and solutions. Based on this conceptual framework, Schwab described three levels of inquiry. At the simplest level, problems and methods are provided to students, and students are requested to discover relations not previously clear from their classroom learning. At a second level, problems are provided to students who are expected to propose their own methods and solutions. At a third level, all three phases of inquiry are left open to students. Derived from Schwab's openness levels of inquiry, Herron (1971) produced a four-point scale that suggests adding a zero level in which problems, methods, and solutions are all provided or immediately obvious from statements in students' task manuals. This four-point scale is frequently used by science educators to distinguish between various levels of inquiry (Banchi & Bell, 2008; Bulunuz, Jarrett, & Martin-Hansen, 2012; Fay, Grove, Towns, & Bretz, 2007; Jiang & McComas, 2015; Lederman, 2009). Some other studies proposed more openness levels by subdividing the previously mentioned three phases of inquiry (Germann et al., 1996; Tamir & Lunetta, 1978). Wenning (2005, 2007) presented a conceptual scheme of eight levels of scientific inquiry, building upon Herron's work, which further clarified the definition of scientific inquiry. There is no doubt that the openness level can be used to characterise inquiry-based tasks in science textbooks; however, the openness level is not the exact response to whether the inquiry-based task is well designed. Because the openness level cannot reflect the function performance of the inquiry-based tasks defined above, these studies cannot judge whether an inquiry-based task is effective.

Studies on the educational function of inquiry in textbooks

The third category of studies on evaluating inquiry-based tasks in science textbooks focuses on the functions the tasks should serve in science teaching and learning, which are intended to measure the quality of inquiry-based tasks. The NRC (2000) created six worksheets to identify quality textbooks, among which Worksheet 2 Analysis of Inquiry as Content analysed whether and how the lessons provided students the opportunity to develop abilities and understandings regarding scientific inquiry. The Biological Science Curriculum Study (BSCS, 2002) developed the Analysing Instructional Materials process comprising two sections, *Paper Screen* and *Implementation/ Pilot*. In the *Paper Screen* section, rubrics for examining science content, the work students do, assessment, and the work teachers do are provided. The rubric for examining the work students do, comprising 28 criteria in four facets, is oriented toward inquiry-based tasks in textbooks. To identify evidence that the abilities and understandings about scientific inquiry are embodied, the four facets include quality learning experience, abilities necessary to conduct science inquiry, understandings about scientific inquiry, and accessibility. No statement regarding understandings about scientific knowledge is included in the analysis inventory, which reflects that both of the above studies treated inquiry-based tasks independently as activities. The function to guide students to conduct and understand scientific inquiry was also considered separately from content understanding in some other studies (Fitzgerald & Byers, 2002; Volkmann & Abell, 2003). Millar (2009) introduced an instrument for evaluating inquiry-based tasks, the Practical Activity Analysis Inventory (PAAI), comprising three components, *Learning Objective*, *Design*, and *Presentation*. The PAAI covers the function that inquiry-based tasks should help students develop understandings about scientific knowledge but omits the function to evaluate whether students develop abilities to conduct scientific inquiry. Thus, each of these studies omitted one or more educational functions of inquiry-based tasks; in addition, new understandings about each function of inquiry-based tasks developed gradually. These instruments could not collect comprehensive and accurate evidence to evaluate the quality of inquiry-based tasks.

The overview of previous research led to the following:

(1) Existing instruments focusing on the instruction of the inquiry process and the openness levels of inquiry can serve the purpose for which they are intended quite well. However, only evaluating the instruction of the inquiry process apparently separates the scientific process from scientific content and omits the other three functions of inquiry-based tasks. The openness level cannot reflect the function performance of an inquiry-based task and cannot address whether an inquiry-based task is well designed.

(2) Although instruments concentrating on functions can provide responses to the quality of inquiry-based tasks, the majority of these instruments, if not all, fail to systematically describe the functions of inquiry-based tasks, which results in an incomplete and inaccurate assessment of the quality of the inquiry-based task. To offer more evidence to make value judgments on inquiry-based tasks, it is necessary to fully consider the functions that inquiry-based tasks in science textbooks should serve. Additionally, because understandings about the functions of inquiry-based tasks have been updated, a new and more valid instrument is required.

(3) Open questions utilise a great deal of space in most tools, which is certainly conducive to gathering evaluators' accurate comments on inquiry-based tasks in textbooks. However, because the inclusion of evaluators' subjectivity is impossible to avoid, users of these tools are expected to be well equipped to understand scientific inquiry and render proper value judgments.

Purpose of the study

Since the early 1950s in Mainland China, the primary and secondary science curriculum has been reformed eight times nationwide. In the latest round of curricular reform, launched at the beginning of 2000, the goal of science education was shifted from the transfer of knowledge to the development of students' scientific literacy with inquiry-based teaching, which placed scientific inquiry at the heart of science education (Liu, Liang, & Liu, 2012). In Mainland China, science is generally taught as an integrated subject in primary schools, either an integrated subject or separate subjects in junior secondary schools and as separate subjects in senior secondary schools. One of the most active areas of curricular reform, secondary biology education has experienced many changes and innovations in recent years (Lu & Liu, 2012). A significant change is the diversified development of biology textbooks, which ended the period in which one single textbook was used nationwide. A great deal of inquiry-based tasks appeared in these textbooks, in accordance with the belief in teaching and learning science by inquiry, prominent in the national biology standards. Perhaps even more than in K-12 classrooms in other areas of the world, textbooks influence what is taught in science classes and how it is taught in Mainland China. Thus, this study was constructed to develop and validate an instrument for measuring the quality of inquiry-based tasks in science textbooks. Based on the previous review, the core research questions addressed in this study are as follows.

How does the instrument developed by this study judge whether inquiry-based tasks are well designed? Does the instrument possess acceptable reliability and validity? Is the instrument reliable and valid for evaluating the function performance of inquiry-based tasks?

Theoretical framework

An instrument – the Inquiry-Based Tasks Analysis Inventory (ITAI) – was developed to gather evidence to evaluate how well inquiry-based tasks in science textbooks perform the above-mentioned functions. Items were developed based on current dominant understandings regarding each function that inquiry-based tasks in science textbooks should serve.

Items to check consistency with curricular knowledge objectives

Inquiry-based tasks in science textbooks should assist students in the construction of understandings about scientific concepts. Since scientific inquiry became fashionable in the mid-1960s, there has been a misconception that inquiry-based learning can occur without attention to scientific knowledge (BSCS, 2005). In fact, students first begin to construct their learning using their prior knowledge of the topic and then inquire into areas that the students do not yet understand, which is also one of the central expectations of IBSE (NRC, 2012; Pratt, 2012). In terms of scientific knowledge, the prevailing attitude is that less is more in K-12 science education. Students are expected to engage in deep exploration of a limited set of fundamental and important concepts rather than the coverage of mile-wide, multiple topics. Core ideas can provide an organisational structure for the acquisition of new knowledge and help prepare students for broader understanding. Furthermore, learning core ideas by scientific inquiry enables students to be less like novices and more like experts (Stewart, Cartier, & Passmore, 2005; NRC, 2012). In light of these perspectives, two items were included in the ITAI scale to check the consistency between inquiry-based tasks and curricular knowledge objectives. Items and the scoring rubric of each item are attached in Appendices 1 and 2.

Items regarding the opportunities for students to use inquiry process skills

To provide students with opportunities to use inquiry process skills is another function that inquiry-based tasks in science textbooks should serve. One rationale for identifying skills in the scientific process is Gagné's Learning Hierarchies Theories, which suggested clarifying students' capabilities of performance in science classes (Gagné, 1963). The American Association for the Advancement of Science (AAAS) launched a series of projects under the title Science – A Process Approach (SAPA) in which 14 process skills appropriate to various scientific disciplines and reflective of scientists' behaviour were identified (Livermore, 1964). SAPA inquiry process skills were widely cited in subsequent studies as a classical and popular method of identification. The National Association for Research in Science Teaching (NARST) reviewed the SAPA inquiry process skills and assigned particular descriptions to 12 of the skills (Padilla, 1990). Wenning (2005) expanded the SAPA hierarchy of process skills by adding two groups of skills, rudimentary skills and advanced skills. This study adopts the SAPA identification and the NARST description of inquiry process skills with the exception of the skill-termed *experimenting* because the description of experimenting overlaps other skills such as formulating hypotheses, controlling variables, defining operations, and interpreting data. Conversely, there is no skill regarding developing questions; therefore, a process skill regarding asking questions replaces the skill of experimenting. A list of inquiry process skills defined by this study is provided in Table 1. All of these skills can be accessed by applying the skills to authentic scientific inquiry. Inquiry-based tasks in science textbooks should provide students opportunities to use these skills. Items of this dimension and a detailed scoring rubric for each item are attached in Appendices 1 and 2.

Items on the reflection of understanding scientific inquiry

The next function of inquiry-based tasks in science textbooks is helping students establish understandings about scientific inquiry, which is as important as having the ability to conduct scientific inquiry (Lederman et al., 2014). Regarding the understandings about scientific inquiry, this study adopted Lederman and Lederman's identification (2012) from the *Second International Handbook of Science Education* and designed eight items regarding this dimension (Appendix 1).

Table 1. List of inquiry process skills.

Skills	Descriptions
Observing[a,b]	Using the senses to gather information about an object or event[b]
Inferring[a,b]	Making an educated guess about an object or event based on previously gathered data or information[b]
Measuring[a,b]	Using standard and nonstandard measures or estimates to describe the dimensions of an object or event[b]
Communicating[a,b]	Using words or graphic symbols to describe an action, object or event[b]
Classifying[a,b]	Grouping or ordering objects or events into categories based on properties or criteria[b]
Predicting[a,b]	Stating the outcome of a future event based on a pattern of evidence[b]
Controlling variables[a,b]	Identifying variables that can affect an experimental outcome, keeping most constant while manipulating only the independent variable[b]
Defining operationally[a,b]	Stating how to measure a variable in an experiment[b]
Formulating hypotheses[a,b]	Stating the expected outcome of an experiment[b]
Interpreting data[a,b]	Organising data and drawing conclusions from it[b]
Asking questions	Raising an appropriate question
Formulating models[a,b]	Creating a mental or physical model of a process or event[b]

[a]SAPA inquiry process skills (Livermore, 1964).
[b]NARST inquiry process skills (Padilla, 1990).

Qualitative analysis of the opportunities to develop HOTS

One more function of inquiry-based tasks in science textbooks is giving students opportunities to develop HOTS. The primary difference between an inquiry-based task and a traditional school activity is that an inquiry-based task requires higher order cognitive processes, such as scientific reasoning and decision-making (Chinn & Malhotra, 2002). In addition, IBSE that promotes HOTS is particularly important in an era in which achieving scientific literacy for all students has become a major goal (Hofstein & Lunetta, 2004; Madhuri, Kantamreddi, & Prakash Goteti, 2012). No consensus identification has yet been reached of HOTS. Several researchers consider the top three cognitive processes in Bloom's Taxonomy of Educational Objectives – analyse, evaluate, and create – to be HOTS (Anderson et al., 2013; Ramirez & Ganaden, 2008), whereas some other scholars derive HOTS from twenty-first century skills such as decision-making, problem-solving, critical thinking, and creative thinking (Conklin, 2012; Heong et al., 2011; Marzano & Pickering, 1997). Respecting that descriptions of HOTS vary from one study to another and that scientific inquiry is not the sole path to developing such thinking skills, this study prefers to consider this dimension in the qualitative discussion section rather than develop items regarding these skills on the ITAI scale.

Thus, the final form of the ITAI scale is a combination of items from the first three dimensions. The instrument ITAI contains a scale comprising 22 items and detailed scoring rubrics for each item. To enhance the objectivity of evaluation results, all items were developed to elicit a yes or no response. The scoring rubrics were developed to justify different responses. A copy of the final version of the ITAI scale and relevant scoring rubrics are offered in Appendices 1 and 2.

Methods and procedures

A pilot study was conducted to regulate the reliability and validity of the instrument.

Sample

The sample included a total of 53 inquiry-based tasks from the three most widely adopted senior secondary biology textbooks in Mainland China. There are many tasks in a textbook, such as assessment tasks, reading tasks, and inquiry-based tasks. This study only considers inquiry-based tasks. According to various descriptions, inquiry refers to the approaches that scientists use, such as investigation and research. Thus, tasks labelled inquiry, investigation, or research in a textbook are considered inquiry-based tasks. The three textbooks, from three different publishers, cover nearly all of the senior secondary biology classrooms in Mainland China. This study labelled the three textbooks as Textbook 1, Textbook 2, and Textbook 3. Consistent with the national senior secondary biology standards in Mainland China, each textbook contains three compulsory modules, Molecule and Cell (Module 1), Heredity and Evolution (Module 2), and Homeostasis and Environment (Module 3), and three optional modules, Application of Biotechnology, Biological Sciences and Society, and Issues in Modern Biotechnology. Because the optional modules are not required, these modules have considerably less influence on classroom teaching and learning. Thus, this study selected the compulsory modules to be the sample. Textbook 1 contains 16 inquiry-based tasks, of which 5 are from Module 1, 3 are from Module 2, and 8 are from Module 3; Textbook 2 contains 12 inquiry-based tasks, of which 5 are from Module 1, 2 are from Module 2, and 5 are from Module 3; Textbook 3 contains 25 inquiry-based tasks, of which 9 are from Module 1, 5 are from Module 2, and 11 are from Module 3. Every inquiry-based task is labelled by a code. The coding rule is that the first number refers to the textbook, the second number designates the module, and the last numbers reflect the page. An overview of the sample composition is listed in Table 2.

Validation of the instrument

In this phase, four science education experts (faculty members in science education) participated in collecting data from the sample (Table 2) using the first version of the ITAI. The four experts are labelled RB, RC, RP, and RW; for each, the first letter R indicates rater and the second letter comes from the initials of the expert's given name. RB is devoted to promoting students' understandings about scientific inquiry and developed a school-based curriculum, An Advanced Course of Scientific Inquiry, in his PhD dissertation. RC developed a solid understanding about scientific inquiry during his study abroad, during which his supervisor was an outstanding expert on the nature of science and scientific inquiry. RP continued RB's work and implemented the inquiry course in a senior secondary school in Beijing. RW is the lead author of this paper. It is reasonable to assume that

Table 2. Overview of sample composition.

	Module 1	Module 2	Module 3
Textbook 1	1161, 1183, 1191, 11,104, 11,110	1257, 1291, 12,116	137, 139, 1351, 1361, 1368, 1375, 1397, 13,102
Textbook 2	2123, 2156, 2164, 2166, 2194	2261, 2273	235, 2360, 2371, 2393, 23,130
Textbook 3	3116, 3123, 3153, 3158, 3167, 3170, 3183, 3187, 3198	3214, 3244, 3251, 3286, 3292	3324, 3338, 3351, 3357, 3365, 3373, 3376, 3387, 3389, 33,105, 33,110

Table 3. Overview of the assignment for raters.

Raters	Module 1	Module 2	Module 3	Inquiry-based Tasks in total
RW	Textbook 1, 2, 3	Textbook 1, 2, 3	Textbook 1, 2, 3	53
RB	Textbook 3	Textbook 1	Textbook 2	17
RC	Textbook 1	Textbook 2	Textbook 3	18
RP	Textbook 2	Textbook 3	Textbook 1	18

all four raters possess a deep understanding about scientific inquiry validating the reliability of the collected data. Before working separately, the four raters gathered in a workshop to discuss the rationale, goals, expectations, and meaning of each item and rubric in this study, resulting in the ITAI's demonstrating high acceptability. Then, the four raters used the ITAI independently. To ensure that evaluation data collected by different raters would fit the same logical scale, assignments in this section are demonstrated in Table 3. In addition, each of the inquiry-based tasks would be reviewed by two raters because of this assignment.

Two pieces of software, SPSS 13.0 for Windows and Winstep for Rasch, were used to analyse the inter-rater reliability, discriminant validity, convergent validity and Cronbach's alpha reliability of the ITAI. Winstep for Rasch is a set of software based on the Rasch model in item reaction theory, which is widely applied in learning progression research for the development of examination items. According to the item fit statistics reported by Rasch software, researchers can determine whether to delete an item. However, Bond and Fox (2007), the developer of the software based on Rasch model, stated that the significance of fit statistics not only determines which item should be deleted from the examination paper but also identifies the misfit item that merits close attention and deeper analysis by researchers. Respecting the advantage of the software based on Rasch model in providing detailed analyses of items and samples (Liu, 2010), Winstep for Rasch was used to examine the reliability and validity of the ITAI in this study. The response 'Yes' was scored 1 and 'No' was scored 0 when inputting data.

Results

Reliability and validity of the ITAI

Inter-rater reliability

Since there are only two response categories, 1 and 0, which is not a string of continuous data, the nonparametric Wilcoxon test was adopted to examine the significance of differences between scores from two different raters who evaluated the same inquiry-based task. The test report is presented in Table 4. As seen in the table, all of the significant coefficient Asymp.Sig. (2-tailed) are much greater than 0.05, which indicates no significant difference between scores from different raters. Additionally, it is a common belief that value judgments rendered by a particular rater always show high internal consistency. RW reviewed all 53 inquiry-based tasks whereas RB, RC, and RP analysed different compositions of the sample separately. Because there was no significant difference between RW's evaluation and RB, RC, and RP's evaluations, it may be inferred that a universal inter-rater consistency exists when using the ITAI. The ratio of identical responses to total responses was also calculated. The statistics (Table 5) indicate that the concordance rate of RW and RB was 89.8%; the rate for RW and RC was 84.1% and the rate for RW and RP was

Table 4. Report of Wilcoxon test on inter-rater difference. All of the significance coefficient Asymp.Sig.(2-tailed) are much greater than 0.05, indicating that there is no significant difference between scores from different raters.

Raters	Sample number	Asymp. Sig. (2-tailed)	Raters	Sample number	Asymp. Sig. (2-tailed)	Raters	Sample number	Asymp. Sig. (2-tailed)
W, B	1257	1.000	W, C	1161	0.317	W, P	137	0.317
W, B	1291	0.157	W, C	1183	0.317	W, P	139	0.564
W, B	12,116	0.157	W, C	1191	0.317	W, P	1351	0.157
W, B	235	1.000	W, C	11,104	0.317	W, P	1361	0.157
W, B	2360	0.414	W, C	11,110	1.000	W, P	1368	1.000
W, B	2371	1.000	W, C	2261	0.180	W, P	1375	0.180
W, B	2393	1.000	W, C	2273	0.102	W, P	1397	0.414
W, B	23,130	1.000	W, C	3324	0.257	W, P	13,102	0.157
W, B	3116	0.655	W, C	3338	0.317	W, P	2123	0.564
W, B	3123	0.317	W, C	3351	0.317	W, P	2156	0.180
W, B	3153	1.000	W, C	3357	0.317	W, P	2164	0.564
W, B	3158	0.102	W, C	3365	0.180	W, P	2166	1.000
W, B	3167	1.000	W, C	3373	0.157	W, P	2194	0.317
W, B	3170	0.180	W, C	3376	0.564	W, P	3214	1.000
W, B	3183	1.000	W, C	3387	0.317	W, P	3244	0.317
W, B	3187	0.317	W, C	3389	0.102	W, P	3251	1.000
W, B	3198	0.317	W, C	33,105	0.564	W, P	3286	0.157
			W, C	33,110	0.564	W, P	3292	1.000

Table 5. Ratio of identical responses to total responses.

Raters	Number of total responses	Number of identical responses	Concordance rate
Compare RW with RB	374	336	89.8%
Compare RW with RC	396	333	84.1%
Compare RW with RP	396	353	89.1%
Total	1166	1022	87.7%

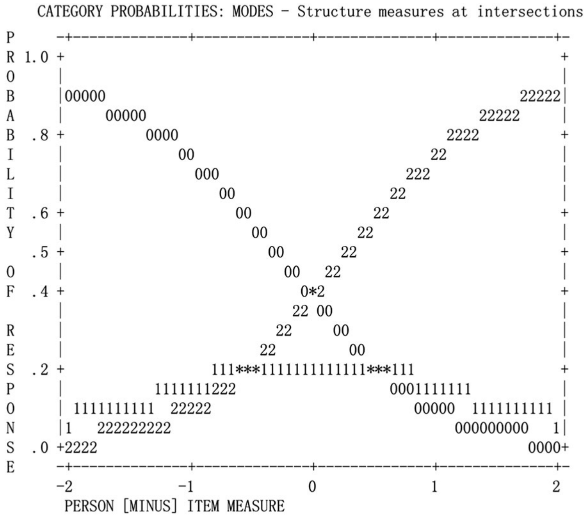
89.1%. The average concordance rate was 87.7%, indicating a high consistency among different raters when using the ITAI to analyse inquiry-based tasks in science textbooks. All of these data indicate that the ITAI has appreciable inter-rater reliability.

Discriminant validity of response categories

Although the ITAI possesses high inter-rater consistency, several different responses exist. With different responses from different raters, this study averaged the two raters' responses for each item as the final score. Consequently, there may be three response categories for each item, 0, 0.5, and 1. In each response category, specific discrimination that does not overlap another response category is absolutely necessary; otherwise, the category must be deleted. The software Winstep for Rasch was used to examine the discriminant validity of the three response categories. Because the software only identifies integers, the three responses 0, 0.5, and 1 were labelled 0, 1, and 2, respectively. An analysis report is provided in Table 6. According to the recommendation of the software guidelines, the minimal number of observed counts per category should be 10, and thresholds should be at least 1.4 apart but not more than 5 to distinguish between categories (Bond & Fox, 2007; Linacre, 1999). The diagnostics presented in Table 6 illustrate that the response category 0.5 does not meet the criteria of all items. In addition, a more visualised distinction among response categories may be embedded in the probability curves (Bond & Fox, 2007). Figure 1 displays the probability curves for the three response categories of Item

Table 6. Analysis report of the discriminant validity of response categories.

Item	Category label	Score	Observed count	Threshold	Item	Category label	Score	Observed count	Threshold
1	0	0	3	NONE	12	0	0	10	NONE
	1	0.5	5	0.55		1	0.5	12	0.15
	2	1	45	−0.55		2	1	31	−0.15
2	0	0	3	NONE	13	0	0	37	NONE
	1	0.5	7	0.19		1	0.5	5	1.19
	2	1	43	−0.19		2	1	11	−1.19
3	0	0	12	NONE	14	0	0	38	NONE
	1	0.5	5	1.20		1	0.5	4	1.47
	2	1	36	−1.20		2	1	11	−1.47
4	0	0	18	NONE	15	0	0	48	NONE
	1	0.5	16	−0.06		1	0.5	2	1.63
	2	1	19	0.06		2	1	3	−1.63
5	0	0	25	NONE	16	0	0	39	NONE
	1	0.5	5	1.38		1	0.5	9	0.27
	2	1	23	−1.38		2	1	5	−0.27
6	0	0	22	NONE	17	0	0	25	NONE
	1	0.5	3	1.92		1	0.5	6	1.38
	2	1	27	−1.92		2	1	22	−1.38
7	0	0	38	NONE	18	0	0	42	NONE
	1	0.5	5	1.23		1	0.5	6	0.72
	2	1	11	−1.23		2	1	5	−0.72
8	0	0	48	NONE	19	0	0	28	NONE
	1	0.5	4	0.38		1	0.5	9	0.66
	2	1	1	−0.38		2	1	16	−0.66
9	0	0	24	NONE	20	0	0	3	NONE
	1	0.5	7	1.00		1	0.5	4	0.78
	2	1	22	−1.00		2	1	46	−0.78
10	0	0	32	NONE	21	0	0	34	NONE
	1	0.5	9	0.58		1	0.5	7	−0.89
	2	1	12	−0.58		2	1	12	0.89
11	0	0	38	NONE	22	0	0	17	NONE
	1	0.5	0	NULL		1	0.5	18	−0.25
	2	1	15	0.00		2	1	18	0.25



```
    CATEGORY PROBABILITIES: MODES - Structure measures at intersections
P     -+--------------+--------------+--------------+--------------+-
R 1.0 +                                                             +
O     |                                                             |
B     |00000                                                   22222|
A     |     00000                                         22222     |
B  .8 +          0000                                 2222          +
I     |              00                             22              |
L     |                000                       222                |
I     |                   00                   22                   |
T  .6 +                     00               22                     +
Y     |                       00           22                       |
   .5 +                         00       22                         +
O     |                           00   22                           |
F  .4 +                             0*2                             +
      |                            22 00                            |
R     |                          22     00                          |
E     |                        22         00                        |
S  .2 +                  111***1111111111111***111                  +
P     |           1111111222                   0001111111           |
O     | 1111111111  22222                         00000  1111111111 |
N     |1   222222222                                   000000000   1|
S  .0 +2222                                                     0000+
E     -+--------------+--------------+--------------+--------------+-
      -2             -1              0              1              2
       PERSON [MINUS] ITEM MEASURE
```

Figure 1. Probability curves for the three response categories of item 1.

1. As shown in the figure, the response category 0.5 which is labelled 1, is completely overshadowed and rendered redundant by the other two categories. All of the probability curves of the other 21 items present nearly the same shape as Item 1. Considering all of these diagnostics, the researchers elected to eliminate response category 0.5.

The four raters met again to review and discuss the items that were scored 0.5 on all inquiry-based tasks. During this discussion, the scoring rubric was fully accepted without argument. The different responses of two raters were not a result of a possible alternative interpretation of the scoring rubric, but a result of clerical error or accidental noncompliance with the rubric. This round of discussion was extremely profitable, facilitating more accurate and objective data. Two different responses for an item ultimately converged in a certain response. Only two response categories, 0 and 1, remained in the following statistical analysis.

Convergent validity of items

There are many methods to analyse the validity of an instrument, such as analysis of test content, response processes, internal structures, relations to other variables, and consequences of testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). All of these methods share a coherent goal of constructing the validity of an instrument, which ensures that the data collected by the instrument are exactly what researchers intended to test. The validity analysis of the software Winstep for Rasch was also established for this purpose (Liu, 2010). In the Rasch modelling approach, both items and persons are located on the same map, which is a logit scale with equal size units. Each item and person is located along the logit scale; the location of more difficult items and more able persons is more positive. If the measure of a person's ability is lower than the measure of an item difficulty, this person is not the logical choice to correct respond to this item; or it is necessary to analyse whether the item is valid or the person's response is unusual (Bond & Fox, 2007). In this study, receiving more 'Yes' answers to an inquiry-based task indicated that the task was of higher quality; fewer 'Yes' responses indicated a more difficult item. An inquiry-based task with lower quality measures than item difficulty measures was unlikely to elicit a 'Yes' response. Consequently, the expectation of item validity in this study was consistent with the expectation in the Rasch modelling approach. Winstep for Rasch was used to analyse the convergent validity of items in the ITAI. Table 7 displays the diagnostics. It is generally recommended that perfect fit statistics would satisfy the following four criteria: (1) Model S.E. < 1, (2) $0.70 <$ Infit-MNSQ $<$ 1.30, $-2.0 <$ Infit-ZSTD < 2.0, (3) $0.70 <$ Outfit-MNSQ < 1.30, $-2.0 <$ Outfit-ZSTD < 2.0, and (4) PTMEA > 0. Table 7 indicates that (1) the Model S.E. of all items is less than 1, which meets the first criterion; (2) most items meet the second criterion, except that the Infit-MNSQ of Item 7 is slightly greater and Item 12 is slightly less, whereas the Infit-ZSTD of Item 7 is slightly greater; (3) the Outfit-MNSQ of Items 5, 7, and 18 is greater than 1.3, Items 10, 12, 13, 15, and 20 are less than 0.7, and the Outfit-ZSTD of Item 18 is greater than 2.0; and (4) the PTMEA of all items except for Item 18 is greater than 0. Thus, eight items, 5, 7, 10, 12, 13, 15, 18, and 20, do not meet the criteria well. In fact, such a result indicates overfit, when an item meets one of the two criteria, namely, (1) Infit-MNSQ < 0.7 and Infit-ZSTD < 0 or (2) Outfit-MNSQ < 0.7 and Outfit-ZSTD < 0. The items fit too well with the validity expectation to be authentic. It

Table 7. Diagnostics of item validity.

		Infit		Outfit		
Item	Model S.E.	MNSQ	ZSTD	MNSQ	ZSTD	PTMEA
1	0.50	1.18	0.6	0.97	0.2	0.37
2	0.50	1.18	0.6	0.97	0.2	0.37
3	0.35	1.02	0.2	0.89	−0.3	0.52
4	0.31	0.96	−0.3	0.90	−0.4	0.51
5	0.31	1.21	1.7	1.51	1.7	0.30
6	0.32	0.85	−1.1	0.72	−1.2	0.60
7	0.35	1.40	2.1	1.99	1.7	0.01
8	0.73	1.05	0.3	0.85	0.2	0.11
9	0.31	1.16	1.2	1.08	0.4	0.39
10	0.34	0.88	−0.8	0.68	−0.6	0.47
11	0.33	0.98	−0.1	1.16	0.5	0.37
12	0.36	0.65	−2.0	0.50	−2.0	0.74
13	0.35	0.76	−1.5	0.55	−0.9	0.53
14	0.35	0.90	−0.6	0.70	−0.5	0.44
15	0.61	1.02	0.2	0.65	−0.1	0.20
16	0.41	1.07	0.4	1.09	0.4	0.23
17	0.31	0.83	−1.4	0.73	−1.0	0.57
18	0.43	1.30	1.1	9.90	5.6	−0.15
19	0.33	0.90	−0.8	0.79	−0.5	0.48
20	0.54	0.76	−0.6	0.57	−0.4	0.56
21	0.33	0.92	−0.6	0.75	−0.6	0.47
22	0.31	0.97	−0.2	0.90	−0.3	0.50

is possible that a few of the subjects trimmed their responses because of surmising the intention of the test. However, in this study, the investigated objects were inquiry-based tasks that were not able to respond to the test items themselves. Raters, as a third party, performed content analysis of inquiry-based tasks in strict accordance with the scoring rubric. There was no possibility of inauthenticity for the five overfit items, 10, 12, 13, 15, and 20. Conversely, the diagnostics of overfit indicated that the scoring rubric of these items was quite precise and valid. Consequently, only three of the above eight items, 5, 7, and 18, were underfit, indicating that the relevant scoring rubrics must be reviewed and revised.

Item 5: ‘In this task, students are expected to use the skill *measuring*, yes or no.’ The initial scoring rubric of this item stated, ‘Please analyse whether students are required to perform measuring as defined in Table 1 to accomplish this task.’ The discussion among the raters demonstrated that weighing chemicals for the preparation of experimental solution frequently appears in inquiry-based tasks. Certainly, weighing belongs to measuring as defined; however, weighing for preparing solution is simply a cookbook approach rather than an inquiry process. Thus, the statement on the scoring rubric was modified to ‘Please analyse whether a measurement of the variables directly related to research questions is requisite for accomplishing this task.’

Item 7: ‘In this task, students are expected to use the skill *classifying*, yes or no.’ The initial scoring rubric of this item stated, ‘Please analyse whether students are required to perform classifying as defined in Table 1 to accomplish this task.’ During the discussion with the other three raters, it was determined that as a result of rater RB’s argument to expand the definition of classifying, the detection and identification of substances (e.g. starch turns blue when interacting with iodine) are regarded as classifying. Broadly speaking, that type of classifying does distinguish containing starch from not containing starch but is not classifying as defined in the inquiry process. Thus, the statement on the scoring rubric was modified to 'Please analyse whether classifying that is rigorously defined in the inquiry process (e.g. biological classification) is requisite for accomplishing this task.'

Item 18: 'The text of this task reflects that *all scientists performing the same procedures may not obtain the same results*, yes or no.' The initial scoring rubric stated. 'If the expected outcome of this task is a foregone conclusion (e.g. the logistic growth model of the population of paramecium lived in a glass container), it could not reflect this term.' The focus of this rubric is that all students may not obtain the same results; however, the premise that all students perform the same procedures to answer the same questions was set aside. In some of the inquiry-based tasks, questions and methods are proposed by students. Although students do not obtain identical results, the students do not perform the same procedures either. Thus the words 'First, please analyse whether the problems and methods are provided by the text of inquiry-based tasks' were added to the statement on the scoring rubric.

All of the inquiry-based tasks re-evaluated the three underfit items, 5, 7, and 18, in accordance with the revised scoring rubrics. Then, Winstep for Rasch was used to examine the convergent validity of the items again (Table 8). Table 8 indicates that there are only two items, 5 and 16, that were not a perfect fit, among which Item 5 met two of the four criteria and Item 16 met three of the four criteria. Bond and Fox (2007) stated that no set of data can have a perfect fit. It is generally accepted that an item meets only three of the four coefficients, Infit-MNSQ, Infit-ZSTD, Outfit-MNSQ, and Outfit-ZSTD. Therefore, Item 16 was fit according to the expectation of validity. The bubble charts (Figures 2 and 3) present a more visualised representation that only Item 5 does not fit the validity expectation well. Each bubble is labelled with a number corresponding to the item. The entire bubble located between −2 and 2 indicates that the item is fit, whereas the bubble centre located more negative than −2 indicates that the item is overfit; if the centre is more positive than 2, the item is underfit. As previously noted, it

Table 8. Diagnostics on validity of items revised.

		Infit		Outfit		
Item	Model S.E.	MNSQ	ZSTD	MNSQ	ZSTD	PTMEA
1	0.53	1.27	0.8	0.92	0.2	0.41
2	0.53	1.27	0.8	0.92	0.2	0.41
3	0.36	1.07	0.4	0.94	−0.1	0.52
4	0.32	1.01	0.1	0.96	−0.1	0.50
5	0.32	1.26	1.9	1.66	2.3	0.31
6	0.33	0.97	−0.2	0.84	−0.7	0.55
7	0.43	1.12	0.5	1.06	0.3	0.20
8	0.73	1.03	0.3	0.67	0.0	0.15
9	0.32	1.25	1.7	1.19	0.9	0.37
10	0.34	0.91	−0.6	0.72	−0.6	0.45
11	0.34	1.00	0.0	1.23	0.70	0.37
12	0.37	0.65	−2.0	0.49	−1.8	0.73
13	0.35	0.74	−1.8	0.53	−1.1	0.53
14	0.35	0.94	−0.3	0.72	−0.5	0.42
15	0.61	0.99	0.1	0.54	−0.2	0.22
16	0.38	1.22	1.1	1.60	1.1	0.17
17	0.32	0.85	−1.1	0.75	−1.0	0.56
18	0.54	1.04	0.2	0.96	0.3	0.19
19	0.33	0.96	−0.3	0.86	−0.3	0.44
20	0.57	0.78	−0.5	0.52	−0.3	0.60
21	0.33	0.94	−0.4	0.76	−0.6	0.47
22	0.32	1.02	0.2	0.97	−0.1	0.49

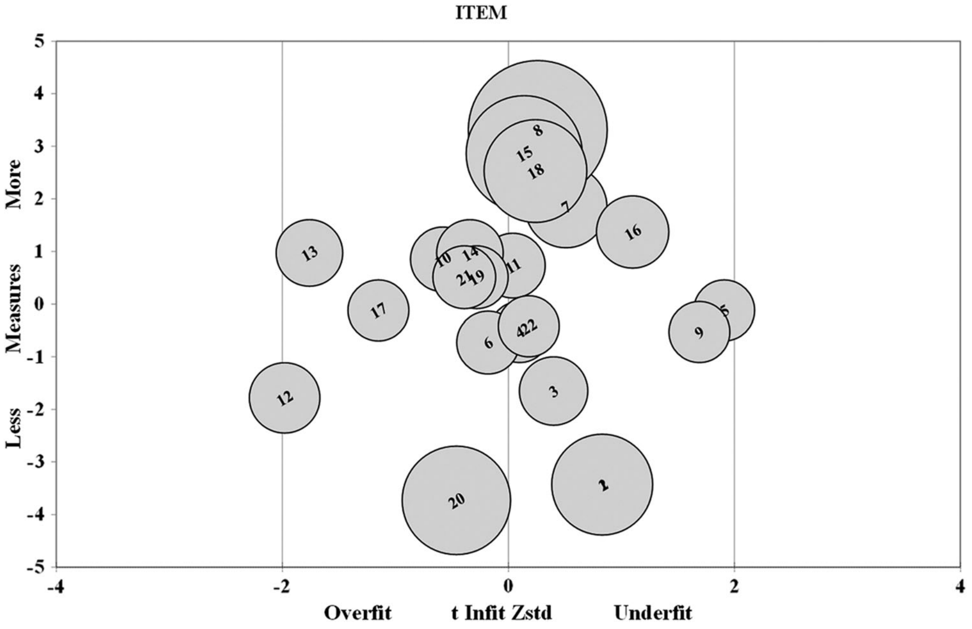


Figure 2. Item fit bubble chart (Infit). One-third of the bubble of item 5 is located positive than 2, but the bubble centre of which falls between −2 and 2.

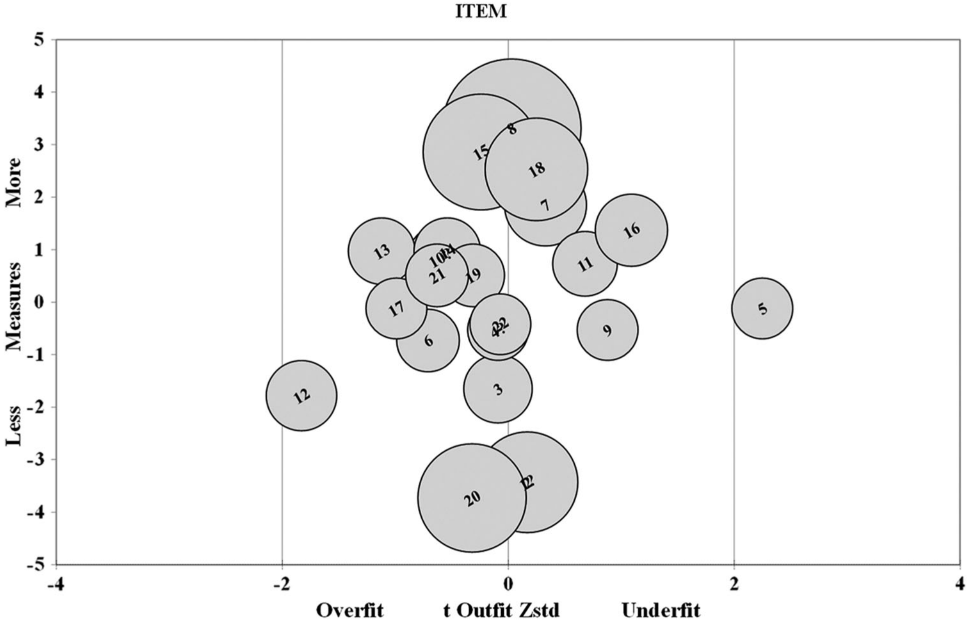


Figure 3. Item fit bubble chart (Outfit). Almost the entire bubble of item 5 falls outside the province of fit, −2~2.

was unnecessary to revise overfit items in this study. In Figure 2, only approximately one-third of the bubble of Item 5 is more positive than 2, whereas the bubble centre falls between −2 and 2. In Figure 3, all bubbles of the items are located in the area of fit,

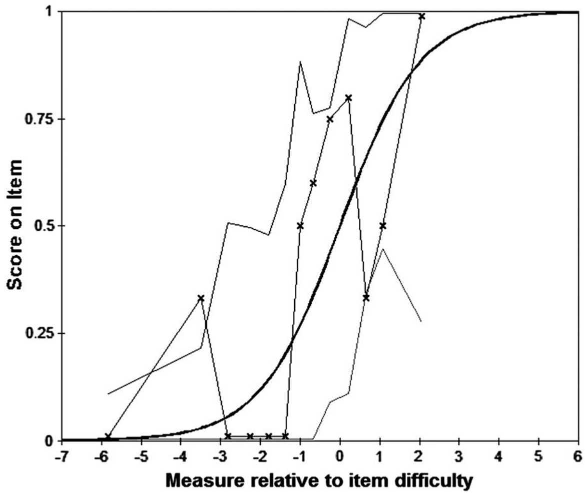
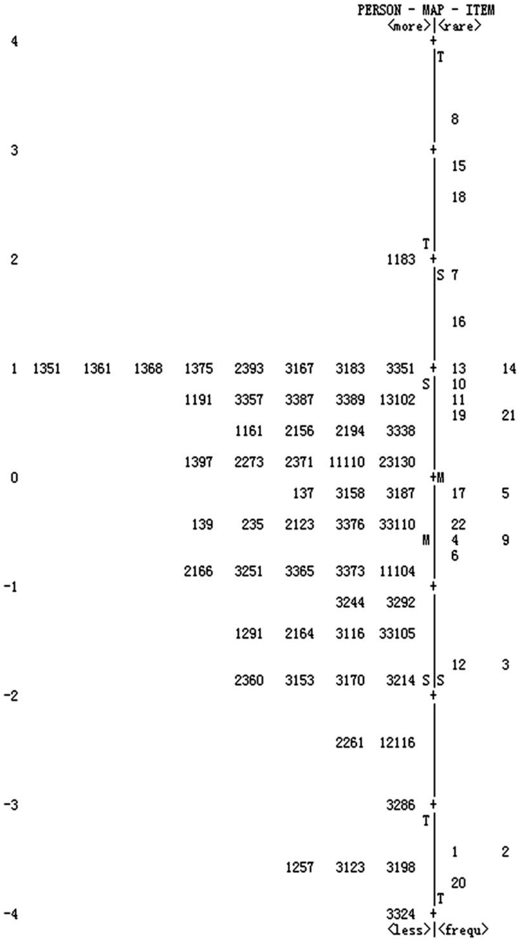


Figure 4. Characteristic curve of item 5. Inquiry-based tasks were grouped into 13 quality levels and represented by ×. The score probability of one group whose quality is 3~4 measures lower than the item difficulty is outside the upper band.

except that nearly the entire bubble of Item 5 falls outside the province of fit. Thus Item 5 is underfit, indicating that the item should be further studied and perhaps omitted. Because Outfit-MNSQs are unweighted mean residuals, some large fit statistics could be because of a few unusual response patterns but not the misfit of items (Liu, 2010). To further identify whether a few individuals responded to Item 5 abnormally, the characteristic curve of Item 5 (Figure 4) was reviewed. In Figure 4, the *x*-axis is the difference between the inquiry-based tasks' quality and the item difficulty, and the *y*-axis is the probability of tasks' receiving a 'Yes' response for this item. For an inquiry-based task, when its quality is higher than the item difficulty, the probability of a 'Yes' response is greater. Similarly, if the quality is lower than the item difficulty, the probability of a 'Yes' response is smaller. The observed pattern is in scatterplots whereas the expected pattern model is in the smooth line, and the acceptable pattern should fall within the band near the expected smooth line. Inquiry-based tasks were grouped into 13 quality levels and are represented by *x* along the *x*-axis in Figure 4. As seen, one group of tasks' probabilities are outside the upper band, indicating that the group of inquiry-based tasks whose quality is 3~4 measures lower than the item difficulty presents an unacceptably high probability of a 'Yes' response for this item. Because the sample size was 53, the size for each group was approximately 4. Based on this result, approximately 4 of 53 individual items were not fit, indicating that the overall fit was good.

To determine exactly which inquiry-based tasks presented unusual response patterns for Item 5, the software Winstep for Rasch was used to draw a Wright Map (Figure 5) on which tasks and items are located along the logit scale with equal size units. The vertical axis is the logit scale, measures of which are on the far left of the Wright Map. Inquiry-based tasks are located along the left side of the axis according to their measures of quality whereas items along the right side of the axis are located according to their difficulty measures. It is not expected that an inquiry-based task located lower than a certain



```
                                          PERSON - MAP - ITEM
                                              <more>|<rare>
4                                                    +
                                                     |T
                                                     |
                                                     |  8
3                                                    +
                                                     |  15
                                                     |  18
                                                     |
                                                    T|
2                                               1183 +
                                                     |S 7
                                                     |
                                                     |  16
                                                     |
1    1351  1361  1368  1375  2393  3167  3183  3351  +  13    14
                                                    S|  10
                         1191  3357  3387  3389 13102 |  11
                                                     |  19    21
                               1161  2156  2194  3338 |
                         1397  2273  2371 11110 23130 |
0                                                    +M
                                      137  3158  3187 |  17    5
                          139   235  2123  3376 33110 |  22
                                                    M|  4     9
                                                     |  6
                         2166  3251  3365  3373 11104 |
-1                                                   +
                                           3244  3292 |
                               1291  2164  3116 33105 |
                                                     |  12    3
                               2360  3153  3170  3214S|S
-2                                                   +
                                                     |
                                           2261 12116 |
                                                     |
-3                                              3286 +
                                                    T|
                                                     |  1     2
                                     1257  3123  3198 |
                                                     |  20
                                                     |T
-4                                              3324 +
                                              <less>|<frequ>
```

Figure 5. The wright map of sample and items. Five inquiry-based tasks 3286, 1257, 3123, 3198, and 3324 located in the area of 3∼4 measures lower than the item 5 difficulty measures.

item could receive an 'Yes' response. In this study, the abnormal individuals appeared in the area of 3∼4 measures lower than the Item 5 difficulty measures. As seen in Figure 5, there were five inquiry-based tasks, 3286, 1257, 3123, 3198, and 3324, located in this area, which were not expected to provide students opportunities to use the inquiry process skill *measuring*. Checking the scores of these tasks on Item 5 revealed that all of these tasks received a score of 0 except for Task 3123, and then Task 3123 was reexamined. In this task, students are expected to use the test paper method to quantitatively test urinary protein. Certainly, a quantitative test of urinary protein fits the definition of measuring in this study; however, this task received a low total score because of several obvious misunderstandings about scientific inquiry. Because of the unusual response pattern of only one inquiry-based task, accounting for approximately 1.89% of the total sample, Item 5 received large fit statistics. Overall, Item 5 fit well and should be accepted without question, which further affirms that the instrument ITAI achieved a good convergent validity after being modified.

Cronbach's alpha reliability

The Cronbach's alpha, the most popular reliability analysis method in current psychological and educational research, was adopted to examine the internal consistency of the ITAI developed by this study. The software SPSS was used to calculate the Cronbach's alpha coefficient. The calculation result indicates that the Cronbach's alpha coefficient was 0.792, close to 0.8, which indicates that the internal consistency of the ITAI was acceptable.

Thus, the ITAI was developed based on the solid foundations of theoretical and empirical argumentation, having high inter-rater reliability, discriminant validity of response category, convergent validity of items, and Cronbach's alpha reliability. Consequently, the ITAI is valid for content analysis of inquiry-based tasks in science textbooks. In addition, the evaluation results gathered by the ITAI are reliable.

Discussion and implications

High-quality inquiry-based tasks in science textbooks are in fact paradigms and practical guides of inquiry-based instruction. In other words, it is the mission of inquiry-based tasks in textbooks to guide teachers and students in performing inquiry-based teaching and learning, for which the quality of inquiry-based tasks is of great concern. Therefore, it is necessary for researchers to pay close attention to the evaluation of inquiry-based tasks in science textbooks.

According to theories of educational evaluation, content analysis of an inquiry-based task in science textbooks is essentially the measurement of how well the task functions as an instructional tool. Previous descriptions of scientific inquiry led to the definitions of four essential functions that inquiry-based tasks in science textbooks should serve. Items and a relevant scoring rubric were created based on current dominant understandings about each function. The instrument ITAI developed in this study into a reliable and valid tool for judging whether inquiry-based tasks perform their functions well after modification and revision. This study extends the previous research on inquiry in textbooks by developing and validating a new instrument that includes a scale and a scoring rubric for content analysis of inquiry-based tasks in science textbooks. The scale and rubric are provided in Appendices 1 and 2.

In follow-up studies, the ITAI was used to evaluate a total of 53 inquiry-based tasks in the three most widely adopted senior secondary biology textbooks in Mainland China, the same sample mentioned in this study. The ITAI provided full and accurate data on the quality of these samples and highlighted the need for revising the current inquiry-based tasks.

The ITAI scale comprises 22 items in three dimensions, and the scoring rubric is a practical guide for grading each item. A higher score on an inquiry-based task indicates that more elements of authentic inquiry are involved in the task, providing a greater challenge for students. The expectation of this study was not that all inquiry-based tasks should achieve high scores, but that an appropriate number of inquiry-based tasks that challenge students' intelligence should be included in one textbook. Additionally, the ITAI is an inventory for content analysis. The purpose of this study is not to predict how teachers and students may conduct inquiry-based tasks; its purpose is to simply review the text of the tasks in science textbooks. Indeed, the function performance of an inquiry-based task depends largely on how teachers and students implement the task in the classroom; however, the quality of the task's content plays a foundational role in supporting its implementation. Data collected by the ITAI are not intended to reflect the implementation effect of an inquiry-based task, simply its content quality. After collecting data by the ITAI, it is strongly recommended that the educational function of developing HOTS be considered in the discussion section.

In addition, the scales and scoring rubrics developed and validated in this study are not only an instrument for evaluating inquiry-based tasks but also principles that should be followed when designing inquiry-based tasks. The ITAI is likely to be used by researchers, textbook evaluators, and teachers to design, assess, and select inquiry-based tasks. Further studies on the reliability and validity of the ITAI may be pursued for a variety of inquiry-based tasks from different disciplines and different regions. More evaluation reports from other people using the ITAI are also necessary. In addition, the ITAI can be used to identify high-quality inquiry-based tasks to provide paradigms for designing and modifying inquiry-based tasks.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

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

Anderson, R. D. (2002). Reforming science teaching: What research says about inquiry. *Journal of Science Teacher Education*, *13*(1), 1–12. doi:10.1023/A:1015171124982

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., … Wittrock, M. C. (2013). *A taxonomy for learning, teaching, and assessing: A revision of bloom's (Pearson new international edition)*. London: Pearson Education.

Banchi, H., & Bell, R. (2008). The many levels of inquiry. *Science and Children*, *46*(2), 26–29. Retrieved from http://www.nsta.org/publications/search_journals.aspx?keyword=VkD55v3bKTg!plus!BtHUn0S5/C84jc29SEFiVbl9Lod9MbY=&journal=QU5/ozhludA=

Biological Science Curriculum Study (BSCS). (2005). *Doing science: The process of scientific inquiry*. Colorado Springs, CO: Author.

Biological Science Curriculum Study (BSCS), & WestEd. (2002). *Analyzing instructional materials (AIM)* (Unpublished work).

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Bulunuz, M., Jarrett, O. S., & Martin-Hansen, L. (2012). Level of inquiry as motivator in an inquiry methods course for preservice elementary teachers. *School Science and Mathematics*, *112*(6), 330–339. doi:10.1111/j.1949-8594.2012.00153.x

Bybee, R. W. (2002). Scientific inquiry, student learning, and the science curriculum. In R. W. Bybee (Ed.), *Learning science and the science of learning: Science educators' essay collection* (pp. 25–36). Arlington, TX: National Science Teachers Association Press.

Bybee, R. W. (2006). Scientific inquiry and science teaching. In L. B. Flick & N. G. Lederman (Eds.), *Scientific inquiry and nature of science: Implications for teaching, learning, and teacher education* (pp. 1–14). Dordrecht: Springer.

Bybee, R. W. (2010). *The teaching of science: 21st century perspectives*. Arlington, TX: National Science Teachers Association Press.

Chiappetta, E. L. (2008). Historical development of teaching science as inquiry. In J. Luft, R. L. Bell, & J. Gess-Newsome (Eds.), *Science as inquiry in the secondary setting* (pp. 21–30). Arlington, TX: National Science Teachers Association Press.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education*, *86*(2), 175–218. doi:10.1002/sce.10001

Conklin, W. (2012). *Higher-order thinking skills: To develop 21st century learners*. Huntington Beach, CA: Shell Educational.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D, Hornik, R. C., Phillips, D. C., … Weiner, S. S. (1980). *Toward reform of program evaluation: Aims, methods, and institutional arrangements*. San Francisco, CA: Jossey-Bass.

Fay, M. E., Grove, N. P., Towns, M. H., & Bretz, S. L. (2007). A rubric to characterize inquiry in the undergraduate chemistry laboratory. *Chemistry Education Research and Practice*, *8*(2), 212–219. Retrieved from http://pubs.rsc.org/en/content/articlepdf/2007/RP/B6RP90031C

Fitzgerald, M. A., & Byers, A. (2002). A rubric for selecting inquiry-based activities. *Science Scope*, *26*(1), 22–25. Retrieved from http://www.nsta.org/publications/search_journals.aspx?keyword=TMLG8m/9FkJwEwY9y!plus!cGxq5KM2Synz7qV2ONz/GT67mbpw6yE73kwbdq0LDSZcSF&journal=rDml7/Vi7wc=

Gagné, R. M. (1963). A psychologist's counsel on curriculum design. *Journal of Research in Science Teaching*, *1*(1), 27–32. doi:10.1002/tea.3660010111

Germann, P. J., Haskins, S., & Auls, S. (1996). Analysis of nine high school biology laboratory manuals: Promoting scientific inquiry. *Journal of Research in Science Teaching*, *33*(5), 475–499. doi:10.1002/(SICI)1098-2736(199605)33:5<475::AID-TEA2>3.0.CO;2-O

Hanauer, D. I., Jacobs-Sera, D., Pedulla, M. L., Cresawn, S. G., Hendrix, R. W., & Hatfull, G. F. (2006). Inquiry learning: Teaching scientific inquiry. *Science*, *314*(5807), 1880–1881. doi:10.1126/science.1136796

Heong, Y. M., Othman, W. B., Yunos, J. B. M., Kiong, T. Z., Hassan, R. B., & Mohamad, M. M. B. (2011). The level of Marzano higher order thinking skills among technical education students. *International Journal of Social Science and Humanity*, *1*(2), 121–125. doi:10.7763/IJSSH.2011.V1.20

Herron, M. D. (1971). The nature of scientific enquiry. *The School Review*, *79*(2), 171–212. Retrieved from http://www.jstor.org/stable/1084259?seq=1#page_scan_tab_contents

Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education*, *88*(1), 28–54. doi:10.1002/sce.10106

InterAcademies Panel (IAP). (2006). *Report of the working group on international collaboration in the evaluation of inquiry-based science education (IBSE) programs*. Retrieved from http://www.ianas.org/meetings_education/files/Santiago_Report_SE.pdf

Jiang, F., & McComas, W. F. (2015). The effects of inquiry teaching on student science achievement and attitudes: Evidence from propensity score analysis of PISA data. *International Journal of Science Education*, *37*(3), 554–576. doi:10.1080/09500693.2014.1000426

Lederman, J. S. (2009). *Teaching scientific inquiry: Exploration, directed, guided, and open-ended levels*. In National geographic science: Best practices in science education. Retrieved from http://www.ngspscience.com/profdev/Monographs/SCL22-0439A_SCI_AM_Lederman_lores.pdf

Lederman, J. S., Lederman, N. G., Bartos, S. A., Bartels, S. L., Meyer, A. A., & Schwartz, R. S. (2014). Meaningful assessment of learners' understandings about scientific inquiry–The views about scientific inquiry (VASI) questionnaire. *Journal of Research in Science Teaching*, *51*(1), 65–83. doi:10.1002/tea.21125

Lederman, N. G., & Lederman, J. S. (2012). Nature of scientific knowledge and scientific inquiry: Building instructional capacity through professional development. In B. J. Fraser, K. G. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education* (pp. 338–339). Dordrecht: Springer Science+Business Media B. V.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*(2), 103–122. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10204322

Liu, X. (2010). *Using and developing measurement instruments in science education: A Rasch modeling approach*. Charlotte, NC: Information Age.

Liu, X., Liang, L. L., & Liu, E. (2012). Science education research in China: Challenges and promises. *International Journal of Science Education*, *34*(13), 1961–1970. doi:10.1080/09500693.2012.709330

Livermore, A. H. (1964). The process approach of the AAAS commission on science education. *Journal of Research in Science Teaching*, *2*(4), 271–282. doi:10.1002/tea.3660020403

Lu, Q., & Liu, E. (2012). Alignment between high school biology curriculum standard and the standardised texts of four provinces in China. *Journal of Biological Education*, *46*(3), 149–164. doi:10.1080/00219266.2011.645855

Madhuri, G. V., Kantamreddi, V. S. S. N., & Prakash Goteti, L. N. S. (2012). Promoting higher order thinking skills using inquiry-based learning. *European Journal of Engineering Education*, *37*(2), 117–123. doi:10.1080/03043797.2012.661701

Marzano, R. J., & Pickering, D. J. (1997). *Dimensions of learning: Trainer's manual (2nd edition)*. Cheltenham, VIC: Hawker Brownlow Education.

McDonald, C. V. (2016). Evaluating junior secondary science textbook usage in Australian schools. *Research in Science Education*, *46*(4), 481–509. doi:10.1007/s11165-015-9468-8

Meyer, D. Z., Meyer, A. A., Nabb, K. A., Connell, M. G., & Avery, L. M. (2013). A theoretical and empirical exploration of intrinsic problems in designing inquiry activities. *Research in Science Education*, *43*(1), 57–76. doi:10.1007/s11165-011-9243-4

Millar, R. (2009). *Analysing practical activities to assess and improve effectiveness: The Practical Activity Analysis Inventory (PAAI)*. York: Centre for Innovation and Research in Science Education, Department of Educational Studies, University of York.

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

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

National Research Council (NRC). (2012). *A framework for k-12 science education: Practices, Crosscutting concepts, and core ideas*. Washington, DC: National Academy Press.

Padilla, M. J. (1990). *The science process skills*. In Research matters – to the science teacher. Retrieved from http://www.narst.org/publications/research/skill.cfm

Pingel, F. (2010). *Guidebook on textbook research and textbook revision (2nd revised and updated edition)*. Pairs: United Nations Educational, Scientific and Cultural Organization.

Pratt, H. (2012). *The NSTA reader's guide to "A framework for k-12 science education: Practices, crosscutting concepts, and core ideas" (Expanded edition)*. Arlington: National Science Teachers Association Press.

Ramirez, R. P. B., & Ganaden, M. S. (2008). Creative activities and students' higher order thinking skills. *Education Quarterly*, *66*(1), 22–33. Retrieved from http://journals.upd.edu.ph/index.php/edq/article/view/1562/1511

Roseman, J. E., Kulm, G., & Shuttleworth, S. (2001). Putting textbooks to the test. *In EisenhowerNational Clearinghouse [ENC] Focus*, *8*(3). Retrieved from http://www.project2061.org/publications/articles/articles/enc.htm

Sax, G., & Newton, J. W. (1997). *Principles of educational and psychological measurement and evaluation* (4th ed.). Belmont, CA: Wadsworth.

Schwab, J. J. (1962). The teaching of science as enquiry. In J. J. Schwab & P. F. Brandwein (Eds.), *The teaching of science* (pp. 1–103). Cambridge, MA: Harvard University Press.

Shulman, L. S., & Tamir, P. (1973). Research on teaching in the natural sciences. In R. M. W. Travers (Ed.), *Second handbook of research on teaching* (pp. 1098–1148). Chicago, IL: Rand McNally.

Stewart, J., Cartier, J. L., & Passmore, C. M. (2005). Developing understanding through model-based inquiry. In M. S. Donovan & J. D. Bransford (Eds.), *How students learn: Science in the classroom* (pp. 515–565). Washington, DC: the National Academies Press.

Tamir, P., & Lunetta, V. N. (1978). An analysis of laboratory inquiries in the BSCS yellow version. *The American Biology Teacher*, *40*(6), 353–357. doi:10.2307/4446267

Tamir, P., & Lunetta, V. N. (1981). Inquiry-related tasks in high school science laboratory handbooks. *Science Education*, *65*(5), 477–484. doi:10.1002/sce.3730650503

Trumbull, D. J., Bonney, R., & Grudens-Schuck, N. (2005). Developing materials to promote inquiry: Lessons learned. *Science Education*, *89*(6), 879–900. doi:10.1002/sce.20081

Tyler, R. W. (1942). General statement on evaluation. *The Journal of Educational Research*, *35*(7), 492–501. doi:10.1080/00220671.1942.10881106

Volkmann, M. J., & Abell, S. K. (2003). Rethinking laboratories: Tools for converting cookbook labs into inquiry. *Science Teacher*, *70*(6), 38–41. Retrieved from http://learningcenter.nsta.org/resource/?id=10.2505/4/tst03_070_06_38

Wenning, C. J. (2005). Levels of inquiry: Hierarchied of pedagogical practices and inquiry processes. *Journal of Physics Teacher Education Online*, *2*(3), 3–11. Retrieved from http://www.physicsfirstmo.org/files/levels_of_inquiry.pdf

Wenning, C. J. (2007). Assessing inquiry skills as a component of scientific literacy. *Journal of Physics Teacher Education Online*, *4*(2), 21–24. Retrieved from http://www2.phy.ilstu.edu/pte/publications/assessing_ScInq.pdf

Appendix 1: Scale of the Inquiry-based Tasks Analysis Inventory (ITAI).

Dimension 1: To assist in the construction of understandings about scientific concepts	Yes	No
1. Scientific concepts involved in this task are consistent with the objectives of the lesson		
2. Understandings about the involved concepts contribute to learning core ideas		
Dimension 2: in this task, students are expected to use the following skills	Yes	No
3.Observing		
4. Inferring		
5. Measuring		
6. Communicating		
7. Classifying		
8. Predicting		
9. Controlling variables		
10. Defining operationally		
11. Formulating hypotheses		
12. Interpreting data		
13. Asking questions		
14. Formulating models		
Dimension 3: The text of this task reflects the following understandings about scientific inquiry	Yes	No
15. Scientific inquiry all begin with a question, but do not necessarily test a hypothesis		
16. There is no single set and sequence of steps or methods followed in all inquiries		
17. Inquiry procedures are guided by the question asked		
18. All scientists performing the same procedures may not get the same results		
19. Inquiry procedures can influence results		
20. Conclusions must be consistent with the data collected		
21. Scientific data are not the same as scientific evidence		
22. Explanations are developed from a combination of collected data and what is already known		

Appendix 2: Scoring rubrics of the Inquiry-based Tasks Analysis Inventory (ITAI)

Before scoring Items 1 and 2, please review the objectives of the lesson and the relevant core ideas in national standards. When scoring Items 3~14, please consult the definition of process skills listed in Table 1. In addition, do not guess how teachers and students may conduct the inquiry-based tasks; simply review the text of the tasks in the science textbooks.

Items	Scoring rubrics
1	If one or more concepts referred to in the objectives of this lesson are applied to accomplish this task or are the conclusions of this task, mark Yes; otherwise, mark No
2	If Item 1 is marked Yes and the concepts involved in Item 1 are components of core ideas described in the national standards, mark Yes; otherwise, mark No
3	If students are required to or must perform observation, mark Yes; otherwise, mark No
4	If students are required to or must infer, mark Yes; otherwise, mark No. Please note the difference between inferring and observing (Lederman & Lederman, 2012)
5	If students are required to or must measure the variables directly related to research questions, mark Yes; otherwise, mark No
6	If students are required to or must communicate as part of this task, mark Yes; otherwise, mark No
7	If students are required to or must perform classifying that is rigorously defined in the inquiry process (e.g. biological classification), mark Yes; otherwise, mark No
8	If students are required to or must predict, mark Yes; otherwise, mark No
9	If students are required to or must control variables, mark Yes; otherwise, mark No
10	If the text of this task completely meets the following three criteria, (1) students are required to define operationally, (2) no cookbook procedure examples are provided in the text, (3) well-defined, scientific, and pertinent research questions are provided or students are asked to develop research questions and no ill-defined, unscientific, and non-pertinent question examples are provided, mark Yes; otherwise, mark No
11	If students are required to or must formulate hypotheses and this task in fact belongs to inquiries that necessarily test hypotheses (e.g. experimental inquiry), mark Yes; otherwise, mark No
12	If students are required to or must interpret data, mark Yes; otherwise, mark No
13	If students are required to ask research questions and no ill-defined, unscientific, or non-pertinent questions or examples are provided, mark Yes; otherwise, mark No
14	If students are required to or must formulate models, mark Yes; otherwise, mark No
15	If the text of this task completely meets the following three criteria, (1) well-defined, scientific, and pertinent research questions are provided or Item 13 is marked Yes, (2) this task does not belong to inquiries that necessarily test hypotheses (i.e. experimental inquiry), (3) students are not required to formulate hypotheses, mark Yes; otherwise, mark No
16	If the task is not designed in accordance with the traditional inquiry template 'asking questions – formulating hypotheses – defining operationally- … ' and students are not required to follow a single set of cookbook steps, mark Yes; otherwise, mark No
17	If the research questions provided are well-defined, scientific, and pertinent or if Item 13 is marked Yes and procedures provided are guided by the questions asked or Item 10 is marked Yes, mark Yes; otherwise, mark No
18	If well-defined, scientific, and pertinent research questions and procedures guided by the questions asked are provided and the expected outcome of this task is not a foregone conclusion (e.g. the logistic growth model of the population of paramecium lived in a glass container), mark Yes; otherwise, mark No
19	If inquiry procedures are open to students' independent design, and no cookbook examples are provided, mark Yes; otherwise, mark No
20	If students are required to reach conclusions based on the data collected, mark Yes; otherwise, mark No
21	If students are required to describe the observations gathered and then analyse and interpret the data, mark Yes; otherwise, mark No
22	If explanations of research questions or questions proposed in the discussion section are necessarily developed from a combination of collected data and what is previously known, mark Yes; otherwise, mark No