The 44th International Association for Educational Assessment (IAEA) Annual Conference

Content Validity of the *Primary Scientific Reasoning Test* – Evidence from the Experts

Ng Yee Ping Diana

Paper was awarded the

Frances M. Ottobre Distinguished Student Scholarship Award 2018

Oxford University Centre for Educational Assessment

University of Oxford

Singapore Examinations and Assessment Board

2018

Abstract

One major goal of science education is the development of scientific thinking and reasoning abilities. Despite significant work conducted on the specific reasoning utilised in science learning, it is unclear how scientific thinking and reasoning is best assessed, or even if existing tests of scientific reasoning adequately measure this higher-order cognitive proficiency. A limited understanding of the nature of scientific reasoning has arguably constrained the development of suitable test instruments. Also, existing tests designed just for primary-aged pupils are few. A new test – the *Primary Scientific Reasoning Test* (PSRT) – was developed and validated with primary school pupils in Singapore to address these key gaps. Original insights from recent science and cognitive science research formed the theoretical basis of scientific reasoning assessed in the test.

This paper reports the results of a study to investigate the content validity of the questions in the initial draft of the PSRT. 18 curriculum and assessment experts from England and Singapore judged the conceptual validity of the scientific reasoning construct, as well as the quality of the questions in the draft of the test. The panel used Sireci's (1998a, 1998b) content validity framework to guide their judgement. A survey with statements on a four-point Likert-like scale collected quantitative ratings from the panel. The survey captured the panel's written comments as well. Using a variety of statistical methods, the analysed feedback provided useful preliminary evidence about the strengths and weaknesses of the questions in the draft test. The paper concludes with a discussion of the advantages and limitations of deploying expert panels to collect content validity evidence from developing and newly constructed tests.

Table of Contents

Introduction

Test content evidence addresses the association between the test content and the underlying construct which the test measures (AERA, APA, & NCME, 2014). Judging the extent to which the assessments tasks are representative samples of the larger domain of performance from the perspective of the test designer (content representativeness) and the users of the test results (content relevance) are common methods to collect validity evidence (Nitko & Brookhart, 2007). Since the 1960s, almost all judgments of content validity involved subject-matter experts (Newton & Shaw, 2014). These experts are selected so that their qualifications, skills and experience match the area of testing (Haladyna & Rodriguez, 2013).

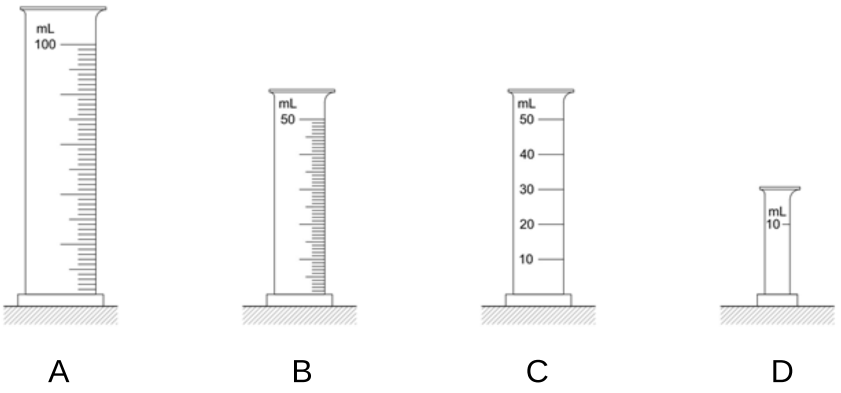
This paper reports the processes and results of a content validity study conducted by a panel of 18 experts on a novel test of scientific reasoning for primary school pupils in Singapore. The study is part of a larger research which developed the scientific reasoning test and investigated its construct validity to address limitations identified in current classroom tests of scientific reasoning. The next section briefly highlights these limitations and the role of the expert panel in the production of content validity evidence.

Background of Study

Rationale

How do primary school pupils reason scientifically? As an illustration, imagine one such pupil attempting the following test item.

The diagram shows four pieces of measuring cylinders.



Which piece will measure 15 cm^3 of water most accurately? Circle your answer below.

A / B / C / D

To get the answer, this pupil will first need to reason that the smallest division on the scale determines the precision of the reading and next determine which equipment gives the most accurate measure. There is involvement of multiple reasoning processes, such as proportional and analogical reasoning, with the content (e.g., concept of volume). Therefore, though the assessment objective is straightforward, what is less clear is how this imagined pupil reasons with the scientific concepts.

As demonstrated with the test item, there is currently a limited understanding of the kinds of performances that exemplify scientific reasoning in science testing. Designing valid and reliable items require well-developed substantive theories about the underlying psychological constructs (Borsboom, 2006). It is thus unclear if and how established learning and measurement frameworks guide the design of many of the existing tests of scientific reasoning (Osborne, 2013). These tests are not validated and lack substantive and psychometric data for score inference (Opitz, Heene, & Fischer, 2017). Understanding the extent of pupil learning and instructional effectiveness become difficult without a basis for valid score interpretation (Pellegrino, DiBello, & Goldman, 2016).

Another limitation of current reasoning tests is the paucity of validated instruments for primary school children. Assessing reasoning problems early and accurately is important as young children intrinsically develop scientific notions from their experiences, and erroneous notions can become resistant to instructional corrections once mentally entrenched (Harlen, 2008). An assessment instrument which produces valid and reliable inferences will inform reasoning flaws and shape rudimentary abilities during this formative period (Harlen, 2007). In brief, major short-comings with current scientific reasoning tests led to the development of a new test – the *Primary Scientific Reasoning Test* (PSRT) for primary-aged pupils. The short discussion below outlines the reasoning construct underlying item design and the validation framework to investigate the construct validity of the PSRT.

Scientific Reasoning Construct and the Mixed-method Validation Framework

Emerging and novel insights across the science research, science education and cognitive science literatures formed the theoretical framework of scientific reasoning underlying the PSRT. Specifically, scientific reasoning is defined as an evaluation of evidence and coordination with theories involving three types of scientific knowledge

– *conceptual*, *procedural* and *epistemic*, while engaged in three science practices to address questions about the ontological, causal, and epistemical aspects of science (Kind, 2013). The practices are *giving scientific explanation*, *designing and evaluating investigations*, and *interpreting and analysing data and evidence*. Subsumed under each practice are three or four sub-practices, which are hypothesised progressions of distinctive skills. For instance, the *giving scientific explanation* practice composes of three sub-practices or skills of modelling abstract ideas, argumentation, and application of knowledge. Guided by this conceptualisation, the items imposed one of three levels of cognitive demand – high, medium, low – while assessing pupils' utilisation of the appropriate knowledge type and science practices. As will be explained later in the *Test Framework* (Materials section of this paper), this conceptualisation steered item design in the PSRT.

Directing the design of the larger research from which this current study derived from is a validation framework drawn from the definitions and operationalisations of validity recommended by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), hereafter referred to as the *Standards.*[1] In this framework, multiple qualitative and quantitative methods systematically collect and analyse evidence from four of the five complementary validity sources advocated by the *Standard*s. Appendix A presents the five-phase framework. Having a variety of data and analyses collected from theoretical perspectives and insiders' views is important when the developed instrument is for the measurement of complex multi-faceted psychological constructs (Onwuegbuzie, Bustamante, & Nelson, 2010). One important *Standards*-recommended source of validity evidence is test content and following established practices from instrument development research, this study collected expert judgement to maximise the content validity of the PSRT (DeVellis, 2012). Sireci's (1998a, 1998b) proposed definition of content validity served as the basis for construct and item review by the experts. The following section discusses the adoption of Sireci's work in the current study.

[1] The *Standards* specifies the criteria for evaluating tests, testing practices, and the effects of test use. According to Zumbo (2014), the recommendations in the *Standards* have come to define the mainstream practices in testing and assessment.

Test Content Validity

Sireci (1998b) defined content validity as “the degree to which a test measures the content domain it purports to measure” (p. 299). Helpfully, Sireci (1998a) ascribed content validity to four test features which jointly guided the gathering of evidence from the experts. Table 1 describes these test features.

Table 1.

Descriptions of Sireci’s four test features.

Test feature	Description
1. Domain definition	observable behaviours indicative of personality trait defined by a psychological construct, and which the items elicit
2. Domain representation	degree to which all the items in the PSRT adequately measure the targeted scientific reasoning construct specified in the table of specifications
3. Domain relevance	alignment between the construct of scientific reasoning and the items
4. Appropriateness of the test development process	measures adopted during the item and test development process to ensure that the test produced valid and reliable inferences

Note. Adapted from Sireci (1998a, 1998b).

On the whole, the expert panel performed nine interrelated, complementary activities and in the process produced an important body of content validity evidence. Table 2 shows how these features relate to the review activities of the experts.

Table 2.

Relationship between Sireci's test features and the review activities.

Test feature	Review	Activity by expert
1. Domain definition	Construct	Relevance, clarity and appropriateness of the domain definition and descriptors
2. Domain representation	Construct-item alignment	Correspondence between the domain definition and descriptors (reasoning practices) to the test items
3. Domain relevance	Cognitive demand	The likely cognitive demand elicited in items matched the intended demand.
4. Appropriateness of the test development process	Item-writing	The design of items against established practices and guidelines.
	Content	Accuracy of scientific content, alignment to the curriculum, as well as accuracy of the assessed knowledge and practices.
	Editorial	Item clarity, technical quality and any grammar, spelling or punctuation errors.
	Sensitivity and fairness	Stereotyping of persons or insensitive use of language.
	Verification of correct answer	Accuracy of intended answers or suggestion of other answers.
	Answer justification	Rationale behind responses from the standpoint of test-takers.

Note. Adapted from Haladyna (2004).

Selection of panel members took into consideration their expertise and knowledge. Outlined next are their profile and professional background.

Expert Panel

The recommended size of an expert panel is between 10 and 20 people to balance between manageability and the gathering of adequate perspectives for robust inferences (Sireci & Faulkner-Bond, 2014). Accordingly, the panel in this study consisted of 18 members, 15 of whom are women. Except for a single member, the experts currently live in Singapore and are familiar with the teaching, learning and assessment of the primary science national curriculum in the country. 15 of those in Singapore come from seven primary schools, while the remaining two are former examiners of the national primary school science examinations in Singapore. The group from schools is composed of teachers and senior curriculum or management personnel, such as a Principal, two Heads of Department and three senior teachers.

The sole member who is not from Singapore is an English examiner and item designer for Key Stage science tests. She was also a former assistant head-master. She was invited to provide expert judgement based on her experience with the English primary curriculum. All members in the panel signed consent forms permitting the reporting of their professional profiles and reviews. At the time of the review, more than half of them were between 30 and 40 years old and five were in their forties. The oldest member was 69 years of age, and the two youngest were in their mid-twenties. Table 3 summarises their teaching and test evaluation experiences.

Table 3.

Professional profile of expert panel.

Profile	Number of experts
Masters or Doctorate degrees in Education	5
Average teaching experience	9.72 years
Test development and evaluation for examinations	
Average experience	9. 67 years
>20 years	2
15-19 years	3
10-14 years	5
5-9 years	4
<5 years	5

Materials

Every member in the panel used three documents to conduct the review. The first was the draft version of the Primary Scientific Reasoning Test (*Draft PSRT)* which had 79 items organised into 32 questions. Each question consisted of one to four items, and had an accompanying characterisation scheme. This scheme described the theoretical knowledge type, science practice and sub-practice from the postulated construct to be assessed by every item in the question. The second document – *Test Framework* – discussed how the theoretical construct of scientific reasoning operationalised into items in the *Draft PSRT*. Part of the discussion included descriptions of the draft construct and the stages of progression of the various sub-practices. Other aspects included the design parameters of the items and the criteria for organising and assembling questions. Finally, the document presented the tables of specifications and exemplar items to illustrate the process of construct-to-item characterisation. See Appendix B for the *Test Framework*.

The final document is the *Proforma*, which is a survey to collect the judgement of the expert panel after reviewing the other two documents. The *Proforma* has a simple bipolar four-point Likert-like scale, with eight statements to gather feedback about the construct and 13 statements about the test items (Oppenheim, 1992). The statements relate to the nine review aspects highlighted in Table 1. The lower two points of the scale indicated disagreement while the higher points signalled agreement ("1" – "strongly disagree", "2" – "disagree", "3" – "agree", "4" – "strongly agree"). Other than rating responses, the *Proforma* also gathered qualitative feedback such as comments and suggestions through the insertion of blank spaces (e.g., after every question). Appendix C shows the statements in the *Proforma* and the specific evidence elicited by each according to Sireci's four features of content validity.

Procedure

This section discusses the preparatory and closure procedures enacted before and after the review, respectively, for four months from June to September 2016. This researcher used SKYPE to conduct meetings with the expert panel in Singapore and London from Oxford, UK. Meetings occurred either on an individual basis or in groups.

Before the review took place in August 2016, this researcher briefed every member at least once about the processes of documentation and review. As the members from the seven schools were unfamiliar with the review procedures of new tests, a preparatory package consisting of the draft construct definition and descriptions, as well as two exemplar items, was sent ahead through email in June 2016. Discussions about the materials in the package took place three to four weeks later in meetings with every school. The discussions largely clarified the definition and construct descriptors, and the alignment between the descriptors and the items. Feedback from the school members during these meetings also helped refine the descriptors.

Subsequently, all members received the test review package consisting of the *Test Framework* with the finalised descriptors, the *Proforma*, and the *Draft Primary Scientific Reasoning Test* (*Draft PSRT*) by email in August 2016. The instructions included the duration of the review period and the procedure (arrangement of online meetings or email) for queries or clarification. There was no issuance of the mark schemes to surface potential problems with the items, such as poor phrasing or unsuitable stimuli. At the end of the one-month review period, members submitted their completed *Proforma* by email. During the review period, there were no queries from the expert panel. After submission, panel members met with the researcher, either individually or in groups. During these closure meetings, which were recorded, reviewers clarified written inputs and shared final thoughts, remarks and ideas. The researcher also took the opportunity to thank them for their participation.

The data collected in the *Proforma* underwent various analyses to organise, summarise and interpret the content validity evidence. Augmenting interpretations of the evidence were transcribed qualitative information collected during closure meetings with the expert panel. A discussion of the analytical methods adopted for inferring the validity from the evidence is in the following section.

Analysis of Data

(a) Rating Scales in *Proforma* about scientific reasoning construct

There are various methods of analysing and summarising data from rating scales in content validity investigations. Sireci and Faulkner-Bond (2014) recommended that besides providing summarised descriptions of domain representation (e.g., table of specifications), there should also be reports of “congruence/alignment statistics” (p. 103). These congruence statistics can come in various forms such as proportion of agreement or outputs from statistical testing. For example, Popham (1992) reported two common ways of summarising content ratings. One way is to compute the majority index, which is the proportion of people which endorsed the survey statement after comparison with the criterion (e.g., alignment to curriculum). The other way is to calculate the arithmetic average of the ratings. Popham added that the means of ratings are more sensitive and conservative measures of approval from the responses of all the reviewers. He recommended that survey statements achieve agreement rates of 75% to 80% to be considered congruent with the criteria.

Other common statistical methods for summarising content ratings included computing the content validity index (Polit & Beck, 2006) and the Aiken’s item content-relevance index (Aiken, 1980). The content validity index (CVI) is a variant of the majority index and is calculated by first dichotomising the four-point nominal scale into “agree” and “disagree”, and then noting the proportion of reviewers rating every statement as acceptable out of the total number of reviewers. Researchers have advocated minimal levels of between 0.8 and 0.9 as indicators of congruity, depending on the purpose of the research (Davis, 1992). Though quick to calculate and easy to understand, this method of collapsing categories into only two also removes potentially relevant information (Polit & Beck, 2006). The Aiken’s index varies between zero and one, and besides indicating the proportion of reviewers rating the survey statements above the midpoint of the rating scale, it can also be used to test specific hypotheses concerning the mean values of the population’s ratings. Unfortunately, the procedure for calculating the Aiken’s index is technically complex (Penfield & Giacobbi, 2004).

Based on the preceding discussion, this study reported the means and the standard deviations of the ratings for each statement about the construct and tabulated the frequency of these ratings. Reporting the content validity indices (CVI) of the ratings related to the construct was not feasible as the collected responses were all positive ("agree" or "strongly agree"). Also carried out was a procedure – the score method – on the mean ratings about the construct to determine their accuracy as estimates of the true population using confidence intervals. Appendix D explains how this method tested the directional hypotheses on the value of the unknown population mean, using the mid-point of 2.5 between the response categories of 1 and 4 as the criterion for assessing the overall endorsement by the panel (Penfield, 2009).

(b) Rating Scales in *Proforma* about items

In contrast to the statistical techniques for treating ratings on the construct, the statements about the items in the *Draft PSR*T reported the content validity indices (CVIs) of ratings. Central tendency measures were redundant as decisions about retaining, amending, or removal of test items required an overall assessment of the item quality. Other information critical for decision-making included item alignment with the curriculum to show that pupils knew the tested content, as well as appropriate technical quality. Also, some types of judgement from the panel had more primacy in these decisions. For instance, doubts cast by a single expert on the accuracy of the science tested in an item could outweigh the collective positive judgment on its relevance to the construct and curriculum.

A final analysis conducted on all ratings was the strategy of data reduction by thematically analysing quantitative data from the responses as a way of "identifying, analysing and reporting patterns (themes) within data" (Braun & Clarke, 2006, p. 79). Accordingly, the analysis and grouping of responses about the construct and the items targeted the production of inferences on the various aspects of content validity. Overall, the treatment of scores from the ratings outlined above mirrored the recommendations of Allen and Seaman (2007) and Carifio and Perla (2007) to take a sensible and appropriate approach in analysis and interpretation. For instance, the means provided a useful way of describing the "typical" opinion of a reviewer but were inadequate for drawing firm conclusions about the underlying construct measured in the rating statement. Table 4 shows the statistics used to summarise the panel members' ratings in the *Proforma*.

Table 4.

Statistics reported on the panel members' ratings.

Reported Descriptive and Quantitative Statistics in *Proforma*	
Ratings about the construct	Ratings about the items
Means Standard deviations Confidence intervals (from score method) Frequency tabulation Thematic analysis	Content validity index Thematic analysis

(c) Qualitative Feedback from *Proforma* and Closure Meetings

Qualitative data in the *Proforma* consisted largely of comments or explications by the panel members about specific survey statements or ratings that they awarded. Together with the transcribed data from the closure meetings, the information provided further understanding of the panel's opinions about the construct and items.

In summary, the study used a variety of data analytical methods to summarise and interpret the expert panel ratings recorded in the *Proforma*. These methods ensured coherent and efficient processing of the ratings as well as greater integration of qualitative and quantitative data for stronger inferences about the content validity of the *Draft PSRT*. The next section discusses the findings and how the panel's inputs clarified the construct and identified problematic items.

Results and Discussion

(a) Content validity evidence of the scientific reasoning construct

Quantitative Evidence - All 18 members rated the eight statements about the scientific reasoning construct positively. More than half of the panel gave strong endorsement to six of the eight statements. The statement which received the strongest agreement was about the relevance of the construct in science. The two statements with the weakest ratings pertained to the clarity of the construct's descriptors and the appropriateness of the proposed progression in reasoning practices. See Figure 1 for a summary of the ratings awarded by the panel.

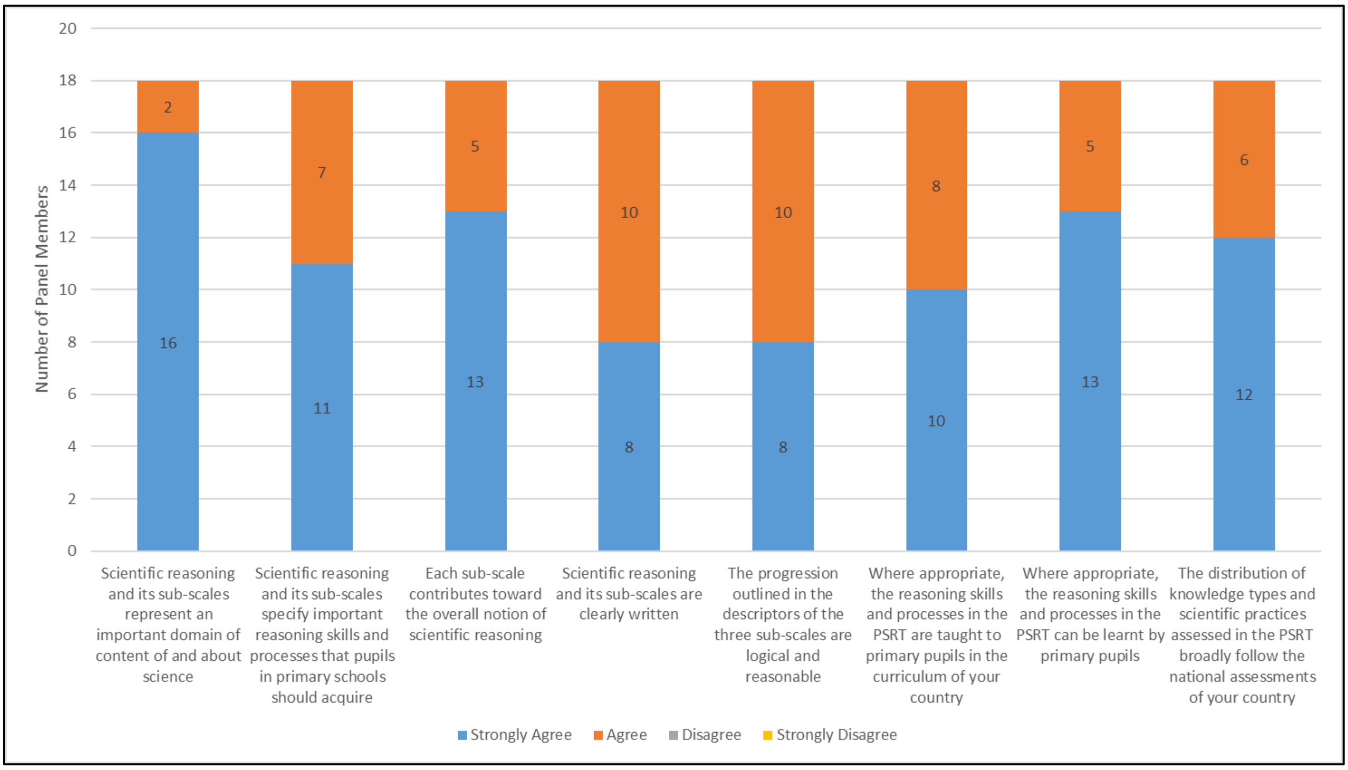


Figure 1. Extent to which members in the panel agreed with the statements about the scientific reasoning construct. Orange bars represent the number of members who agreed. Blue bars represent the number of members who strongly agreed.

The mean and standard deviation showed the smallest spread of scores in the first statement, which had the best agreement, and the widest spread in the two statements with the worse endorsement. See Table 5 for the means, standard deviations and confidence intervals of the ratings. As mentioned earlier, the study adopted a criterion-based validity test using hypothesis testing. From the confidence intervals shown in the same table, there was rejection of the null hypothesis that the population mean μ = 2.5 against the alternative directional hyothesis that μ > 2.5 using the specified Type 1 error rate of α = 0.05. The rejection came about because the lower limits of the 95% confidence intervals of the ratings were greater than the null hypothesis value of 2.5. Therefore, it is argued in this paper that there was valid favourable endorsement on all eight statements.

Table 5.

Means, standard deviations, and 95% score confidence intervals for eight statements about the scientific reasoning construct.

	Statement	**Mean**	**SD**	**95% Confidence Index**	
				Lower Limit	**Upper Limit**
1	Scientific reasoning and its sub-scales (knowledge type, practice and sub-practice) represent an important domain of content *of* and *about* science.	3.89	0.32	3.44	4.34
2	Scientific reasoning and its sub-scales specify important reasoning skills and processes that pupils in primary schools should acquire.	3.61	0.50	3.15	4.07
3	Each sub-scale contributes toward the overall notion of scientific reasoning.	3.72	0.46	3.26	4.18
4	Scientific reasoning and its sub-scales are clearly written.	3.44	0.51	2.98	3.91
5	The developmental progression outlined in the descriptors of each of the three practices is logical and reasonable.	3.44	0.51	2.98	3.91
6	Where appropriate, the reasoning skills and processes in the PSRT are taught to primary pupils in the curriculum of your country.	3.56	0.51	3.09	4.02
7	Where appropriate, the reasoning skills and processes in the PSRT can be learnt by primary pupils.	3.72	0.46	3.26	4.18
8	The distribution of knowledge types and scientific practices assessed in the PSRT broadly follow the national assessments of your country.	3.67	0.49	3.21	4.13

Qualitative Evidence - Feedback from the panel further elucidated opinions about the construct. Unsurprisingly, Statement 1 garnered just two comments affirming the relevance of the construct in science. Statement 2 received comments representing divergent views; two teacher members opined that some of the scientific reasoning knowledge and practices cannot be attained by academically weaker pupils or by primary pupils even, while the Principal expressed that the existing curriculum covered substantial aspects of the knowledge and practices. There were no comments on the third statement.

Statement 4, which sought to elicit views about the clarity of the descriptors, received the most remarks and suggestions. Though there was broad agreement that the descriptors were clear, members highlighted specific descriptors which they found difficult to interpret. For instance, four members identified the descriptors of some sub-practices under the broader practice of *giving scientific explanation,* as being vague. One of two other commonalities among the feedback from the panel about this statement was the difficulty in distinguishing between procedural and epistemic knowledge. One reviewer attributed the confusion to the wording in the *Test Framework* document. The reviewers also commented on the formatting and presentation of the descriptors of the sub-practices, which were considered lengthy.

Statement 5 elicited concerns about the reasonableness of the proposed progression of manifested behaviours as pupils advanced in their capabilities in the various reasoning sub-practices. While the comments on statement 5 were overall positive and indicated broad agreement, many reviewers expressed the need for greater specificity and clearer delineation between stages in some of the sub-practices. As one reviewer explained, such enhancements would “improve coherence and usability”.

Feedback to statement 6 contained opinions with a variety of perspectives. A significant number opined that the curriculum and its learning outcomes did not outline the reasoning, knowledge and practices espoused in the *Draft PSRT*, and pupils were instead implicitly taught. A few added that these abilities were not consistently taught across Singapore schools. The English reviewer highlighted that the teacher often serves as a key factor in the acquisition of these abilities. She explained that, “there may still be the desire to ensure content is covered and this is then at the expense of the skills and reasoning.” In sum, feedback from the Singapore reviewers expressed a preference for the abilities to be articulated more clearly in the existing curriculum. As another reviewer explained, this articulation could help teachers impact “pupils’ capability to self-regulate and learn independently”. Interestingly, the lack of clear guidelines in Singapore’s primary curriculum about types of knowledge and practices in use resonated with comments from the last statement, as will be described shortly.

Feedback to statement 7 acknowledged that primary pupils could learn the knowledge and practices in the *Draft PSRT* given the right contexts and opportunities. However, one reviewer said that the teaching and learning of epistemic knowledge might prove challenging, as many teachers and pupils were “fixated on ‘one right answer’ to questions”. For this reason, he approved the relatively low proportion of questions assessing epistemic knowledge in the *Draft PSRT*.

The eighth and last statement collected opinions about the mark weighting between the subscales in the *Draft PSRT* and the national assessments. Interestingly, there were opposing views, and these came from the English reviewer and a Singapore panel member. The former agreed that there is general alignment between the distribution of knowledge type and practices assessed in the *Draft PSRT* and England’s national assessments. The English reviewer elaborated that, “the frameworks in primary school weave the procedural and epistemic through the content. There is a mixture of low cognitive demand of some content and the higher cognitive demand is often where the epistemic knowledge types are assessed.” In contrast, the Singapore panel member felt that the national assessments in Singapore rarely tested epistemic understanding. He also opined that there was limited assessment of procedural and experimental design. However, his judgement that there was a misalignment seemed isolated, as another member opined that the *Draft PSRT* questions were modelled largely after the style of Singapore’s national examinations.

Cross-over Analysis - Augmented by the panel’s qualitative insights about the construct, further analysis framed the quantitative ratings thematically to Sireci’s (1998a) four test features of content validity. Comments about the construct addressed two of the four features – domain description and appropriateness of the test development processes. The first five statements addressed the relevance, completeness, appropriateness and clarity of the domain description on scientific reasoning. It was clear that all members concurred the construct to be relevant to the learning of science and the primary science curriculum. They also generally agreed that the descriptors of the knowledge types and reasoning practices were comprehensive, helpful and explicit. There was consensus as well that the proposed progression of the reasoning practices was logical and reasonable. However, the relatively lower ratings and comments on statements 4 and 5 indicated that some of the descriptors needed modification or reconsideration. Reviewers cited conflicting terminology, discrepancies between the descriptors and their corresponding examples, as well as conceptual overlaps between descriptions as reasons for amendments.

Statements 6 and 7 solicited views about opportunities for school exposure, as well as the age-appropriateness of the knowledge types and practices outlined in the domain description. Lower ratings and the accompanied qualitative feedback from the Singapore members suggested variability of opportunities across schools that depended heavily on the teachers' experience and knowledge. Overall, there was agreement that primary pupils could learn the espoused knowledge and practices.

The eighth statement checked opinions about the weighting of knowledge types and practices between the *Draft PSRT* and the national assessments. Their ratings provided validity evidence for the appropriateness of the testing approach. Feedback was somewhat mixed. Other than a single comment that the national assessments in Singapore seldom assessed epistemic knowledge, the ratings and feedback from the remaining panel members suggested that there was broad alignment. The next section reports the panel's views about the individual items in the *Draft PSRT*, which produced further validity evidence about the domain representation, domain relevance and the appropriateness of the testing approach.

(b) Content validity evidence of the items in the *Draft PSRT*

Quantitative Evidence - Panel members evaluated each of the 79 draft items on 13 criteria or statements (see Appendix C for statements). Calculation of the content validity indices (CVI) began by collapsing the 4-point ratings into two categories of "agree" and "disagree", and then noting the proportion of reviewers who agree to each statement about the item. [2] In a practice aligned with Popham's (1992) recommendation to consider CVIs above 80%[3] as attainment of criteria congruency,

[2] For brevity, the CVIs of the 79 items are not presented in this paper.
[3] As a gauge, CVIs of 0.83 and below in this study meant that 3 or more panel members disagree.

there was flagging of ratings below 0.83 for attention. Analysis of the CVIs across the items showed that the focus of the panel's disagreements centred on the extent of correspondence between the targeted construct of scientific reasoning in the item characterisation schemes and the test items. Specifically, reviewers disagreed with the assigned subscales, comprising knowledge types and practices, and the cognitive demand of the items. In all, 22 items received a CVI of below 0.83 in at least one rating.

Domain representation is the extent of correspondence between *all* the items on the PSRT and the targeted aspect of the construct (Sireci & Faulkner-Bond, 2014). Following the example shown in Sireci (1998b), the mean CVIs from all the items in statements 4 and 5 are the indices of domain representation for the subscales of knowledge types and practices and cognitive demand respectively. As shown in Tables 6 and 7, the obtained indices were 0.93 and 0.91, which meant that on average for every item, about 17 out of 18 panel members agreed that the subscales and the cognitive demand aligned to the targeted construct. Both indices were above the congruency criteria recommended by Popham (1992) and Davis (1992).

Table 6. *Number of items by CVI levels and the mean CVI obtained to statement 4.*

	Content Validity Index of 79 Items on Statement 4 (subscales)							**Mean CVI**
	1	0.94	0.89	0.83	0.78	0.67	0.56	0.93
No. of items	40	19	4	3	10	2	1	

Note. Items with CVI at and above 0.89 indicated an acceptable level of alignment to the targeted construct (shaded in grey).

Table 7. *Number of items by CVI levels and the mean CVI obtained to statement 5.*

	Content Validity Index of 79 Items on Statement 5 (cognitive demand)									**Mean CVI**
	1	0.94	0.89	0.83	0.78	0.72	0.61	0.56	0.39	0.91
No. of items	34	20	6	8	4	3	2	1	1	

Note. Items with CVI at and above 0.89 indicated an acceptable level of alignment to the targeted construct (shaded in grey).

Domain relevance addresses the degree to which *each* item on the PSRT matches the descriptions in the table of specifications (Sireci & Faulkner-Bond, 2014). This study targeted a pre-defined level of around 0.9 to evaluate the relevance of individual items. This higher standard surfaced more items for scrutiny at this preliminary stage of instrument development. However, striving for 100% agreement might be excessively demanding because there were many reviewers in this initial validation study on a new scientific reasoning construct and disagreements were therefore expected (Polit & Beck, 2006). Based on the designated acceptable level, the proportion of items out of the total with CVIs at and above 0.89 served as indices of domain relevance for subscales and cognitive demand respectively. Therefore, as shown in Tables 6 and 7, the indices of domain relevance were 0.80 for subscales and 0.76 for cognitive demand. These indices meant that only 80% and 76% of the 79 items achieved acceptable levels of alignment to the sub-scales (knowledge type and practices) and cognitive demand. The lowered indices were unsurprising as members rated many items on statements 4 and 5 less favourably. Presented next is a summary of the key issues gathered from the members' qualitative inputs from the *Proforma* and the closure meetings.

Qualitative Evidence – Most of the comments in the *Proforma* were elaborations about poor ratings on the items. Based on how the panel rated the items, it was unsurprising that these remarks centred on the characterisation of the sub-scales and cognitive demand. Other concerns reflected in the *Proforma* included accuracy of the tested science content, clarity of phrasing and choice of vocabulary. These same issues surfaced during the closure meetings, along with other matters such as choice of the most suitable item format for examining pupils' reasoning, and the implications for pedagogy and assessment practices. In summary, qualitative evidence from the *Proforma* and closure meetings substantiated the ratings awarded by the panel. The next section discusses the thematic findings from a cross-over analysis conducted on the quantitative ratings of the items.

Cross-over Analysis – The quantitative ratings addressed three of Sireci's (1998a) four test features of content validity – domain representation, domain relevance and appropriateness of the test development process. Ratings from six of the ratings (Statements 1, 2, 3, 6, 7 and 8) elicited evidence about the development process. Specifically, the first three statements checked the relevance, accuracy, and curricular alignment of the scientific concepts assessed in each *Draft PSRT* item. Most of the items received endorsement for the relevancy of the scientific concepts tested. There were instances where some reviewers expressed doubts over the accuracy and the curricular alignment of the concepts. No more than two reviewers indicated disagreements to each of the first three statements about the scientific concepts. Likewise, for Statements 6, 7 and 8, which pertained to the technical quality of the items, disagreements did not exceed more than three reviewers for identified items.

Statements 4 and 5 collected evidence of domain representation and domain relevance. As discussed earlier, the indices of domain representation suggested that on average for every item, 17 out of the 18 panel members agreed that its subscales and cognitive demand aligned to the suggested characterisation schemes in the table of specifications. However, based on the CVI criteria set for acceptable levels of alignment to the sub-scales (knowledge type and practices) and cognitive demand, only 80% and 76% of the 79 items achieved these levels respectively. From the overall evidence presented in the findings thus far, there were two probable reasons why higher levels of domain relevance were not achieved in this *Draft PSRT*. Firstly, evidence about the appropriateness and clarity of the construct suggested there were issues with the domain descriptions which had, in turn, imposed alignment issues for the reviewers at the item level. The second area where misalignment between item and targeted construct could have occurred was at the item level. Items were either poorly designed or assessed inaccurate or out-of-curriculum scientific concepts, knowledge types or practices.

Conclusions

Overall, this paper presented content validity evidence of the *Draft PSRT* from the expert panel. Analysis of the evidence took place using Sireci's four test features of content validity. From the angle of the domain description of the scientific reasoning construct, the expert panel strongly endorsed its relevance to science learners and the primary science curriculum. The expert panel also generally agreed that the detailed domain description of the knowledge types and reasoning practices were wide-ranging, useful, explicit and could be learnt by primary pupils. However, the panel highlighted that some of the descriptors required modification or reconsideration. The evidence collected from the ratings for domain representation suggested adequate representation for the subscales and the cognitive demand by the items. Though necessary at this stage of instrument development, establishing a high index of content validity as a criterion had reduced domain relevance for individual items.

Inferences about the appropriateness of the testing approach came from three types of validity evidence in this study. First, there was some evidence of appropriate alignment between the weighting of knowledge types and practices to the national assessments. This evidence is important because of the production of stronger inferences from pupils' performances on the *Draft PSRT* given their familiarity with the format and content of the national assessments. Second, evidence from the panel suggested that the scientific concepts assessed in most of the *Draft PSRT* items were relevant, accurate, and aligned to the science curriculum. Third, the technical quality of the items received general endorsement from the panel.

While the inferences made from the panel's feedback had been tremendously useful, they did not each hold equal primacy in final decisions about item retention, modification, or removal from this *Draft PSRT*. Nor were inferences from this study the only evidence of content validity to be considered for these decisions. Evidence from other methods were necessary during the development process to corroborate or challenge interpretations about the robustness of the instrument. To this end, there was collection of data from other studies *s*uch as the psychometric information of pupils' performance and their feedback from cognitive interviews. Results of these studies will be reported elsewhere.

In conclusion, this study illustrates the procedures involved in gathering content validity evidence from an expert panel based on modern validity theory. Overall, the panel provided valuable content validity evidence for making inferences about the strengths and weaknesses of the *Draft PSRT* and its items. However, collecting such evidence required advanced planning, which was time-consuming, laborious and resource intensive. Also, there was a need to weigh and cogently balance among the individual feedback (quantitative and qualitative) from multiple members by considering the salient merits and implications of each. Finally, it is crucial to select simple, easily interpretable, and yet robust statistical items indices. Being cognizant of the advantages and limitations of convening expert panels for collecting content validity evidence ensures that inferences are relevant and contribute to the construct validity of a new instrument.

References

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement, 40*, 955-959.

Allen, E. I., & Seaman, C. A. (2007). Likert scales and data analyses. *Quality Progress, 40*(7), 64-65.

Borsboom, D. (2006). The attack of the psychometricans. *Psychometrika, 71*, 425-440.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*, 77-101.

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences, 3*(3), 106-116.

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5*, 194-197.

DeVellis, R. F. (2012). Scale development: Theory and applications.

Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Harlen, W. (2008). Science as a key component of the primary curriculum: a rationale with policy implications. *Perspectives on Education (Primary Science), 1*, 4-18.

Johnson, R., & Onwuegbuzie, A. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*, 14-26.

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

Morse, J. M. (1991). Approaches to qualitative-quantitative methodological triangulation. *Nursing Research, 40*(1), 120-123.

Newton, P. E., & Shaw, S. (2014). *Validity in Educational and Psychological Assessment*. London: SAGE Publications Ltd.

Onwuegbuzie, A. J., Bustamante, R. M., & Nelson, J. A. (2010). Mixed research as a tool for developing quantitative instruments. *Journal of Mixed Methods Research, 4*, 56-78.

Opitz, A., Heene, M., & Fischer, F. (2017). Measuring scientific reasoning - A review of test instruments. *Educational Research and Evaluation*, 1-24.

Oppenheim, A. N. (1992). *Questionnaire design and attitude measurement*. London: Continuum.

Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity, 10*, 265-279.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*, 59-81.

Penfield, R. D., & Giacobbi, P. R. J. (2004). Applying a score confidence interval to Aiken's item content-relevance index. *Measurement in Physical Education and Exercise Science, 8*(4), 213-225.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in nursing & health, 29*, 489-497.

Popham, W. J. (1992). Appropriate expectations for content judgements regarding teacher licensure tests. *Applied Measurement in Education, 5*, 285-301.

Sireci, S. G. (1998a). The construct of content validity. *Social Indicators Research, 45*, 83-117.

Sireci, S. G. (1998b). Gathering and analyzing content validity data. *Educational Assessment, 5*, 299-321.

Sireci, S. G., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26*, 100-107.

Zumbo, B. D. (2014). What role does, and should, the test standards play outside of the United States of America? *Educational Measurement: Issues and Practice, 33*, 31-33.

Appendices

Appendix A Summary of Mixed-methods Validation Framework

Phase of research, research sequence and validation goal	Mixed-method focus in phase*	Method	Analytic technique	Cross-over analysis
Phase I: Initial Test Design To collect evidence based on test content ↓ *Draft PSRT* ↓ ↓	QUAL	• Review of substantive theories (conceptual, measurement and test design)	• Nomological network analyses	
Phase II: Expert Panel Review To collect evidence based on test content **Phase III: Pilot Study** To collect evidence based on test content and response processes ↓ *Intermediate PSRT* ↓	QUAN + qual (phase II) QUAN + QUAL (phase III)	• Review by expert panel • Field-test of *Draft PSRT* • Conduct cognitive interviews	• Survey analyses • Descriptive analyses • Classical test theory • Content and thematic analyses of participants' responses (instrument and interviews)	• Integrated data reduction (phase II) • Data integration (phases II and III) • Data comparison (phases II and III)
Phase IV: Expert Panel Review (2nd round) To collect evidence based on test content ↓ *Final PSRT* ↓	QUAL	• 2nd review round by expert panel	• Analysis of reviews	• Data comparison (phases II, III and IV)
Phase V: Main Study To collect evidence based on test content, response processes, internal structure & relations to other variables	QUAN + qual	• Administer *Final PSRT* • Administer two psychological instruments - Raven's Standard Progressive Matrices – Plus and the Mill-Hill Vocabulary Scale • Conduct cognitive interviews	• Descriptive analyses; • Rasch modelling; • Correlation analyses; • Structural equation modelling; • Content and thematic analyses	• Data comparison • Warranted assertion analyses • Integrated data display

Note:*Adopting the notation of Johnson and Onwuegbuzie (2004) and Morse (1991), "QUAL" refers to qualitative, "QUAN" to quantitative, "+" stands for concurrent. The use of capital letters refers to the priority or weight of the methodological orientation. For instance, "QUAL + QUAN" denotes both orientations having equal priority in the research phase.

Appendix B *Test Framework*

The *Test Framework* describes the considerations involved in the assembly of the initial items into the *Draft PSRT*. It begins with the item characterisation approach to align the assessment objectives of the items to the scientific reasoning construct. The remaining sections of this appendix discuss the organisation of items and the criteria for mark distribution.

Characterisation of Items

In the *Draft PSRT*, items characterised by one of three levels of cognitive demand measured scientific reasoning by assessing pupils' scientific understanding in implementing specific practices. Figure 1 shows a three-dimensional grid-model representing the knowledge, skills and depth of assessment.

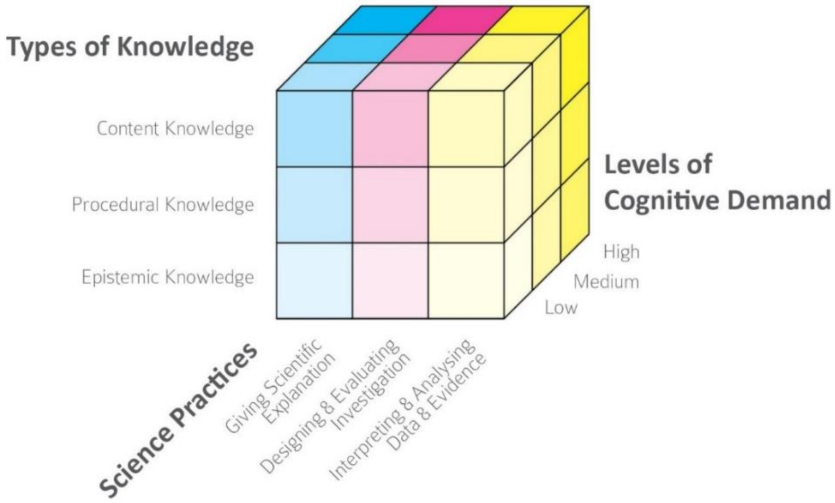

Types of Knowledge
Content Knowledge
Procedural Knowledge
Epistemic Knowledge
Levels of Cognitive Demand
High
Medium
Low
Science Practices
Giving Scientific Explanation
Designing & Evaluating Investigation
Interpreting & Analysing Data & Evidence

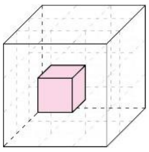
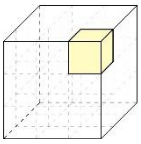
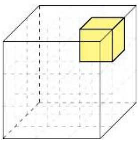
Figure 1. A grid-model for characterising items along the axes of knowledge type, science practices and levels of cognitive demand. Each of the three colours signify a science practice.

The 27 smaller cubes in Figure 1 represent the range of possible item characterisation. In the figure, cubes of the same colour share a common science practice, and the intensity of the colour relates to higher demand. For instance, the darkest yellow cube at the bottom right-hand corner represents an item of high cognitive demand which measures epistemic understanding in the practice of interpreting and analysing data and evidence.

As shown in Table 1, the grid-model forms the basis of the characterisation scheme of three items in a question (see Figure 2) of the *Draft PSRT*. The scheme identifies the specific knowledge, practice, cognitive demand, mark allocation, and item format associated with each item. There is identification of the related sub-practice as well. Although not elaborated in this appendix, the descriptors in each sub-practice outline progressions of increasingly developed and sophisticated attainment of skills associated in the practice. Finally, the scheme also provides a visual placement of the item within the grid-model. The next section elaborates on the arrangement of items in the instrument for administration, and the final number and type of questions for inclusion in the test. Appendix B concludes with the rationale for proportioning and assembling items to reflect the assessment focus in a table of specifications.

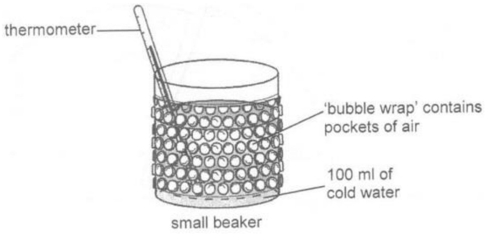
Table 1.

Characterisation scheme of a three-item question in Figure 2.

Characterisation scheme			
	Item 1	Item 2	Item 3
Knowledge type	Procedural	Content	Content
Practice*	2	3	3
Sub-practice	2.1.1	3.3.2	3.3.3
Cognitive Demand	Low	Low	Medium
Mark	1	1	1
Item format	SROS	SROS	CRSS
Grid-model placement			

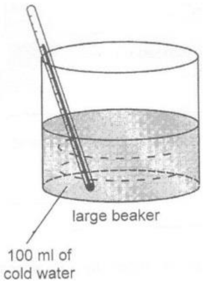
Note. *Practice 1 – Giving scientific explanations, Practice 2 – Designing and evaluating investigative approaches, Practice 3 – Interpreting and analysing data and evidence

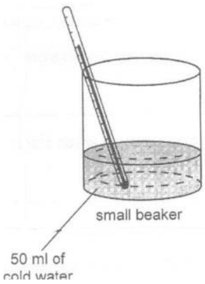
John wants to study whether the material, 'bubble wrap,' affects how cold water gains heat from the room. The diagram below shows one set-up of his experiment.

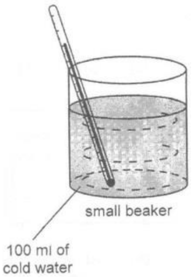


Item 1: Which other set-up shown below, P, Q, or R, must he use in his experiment? Choose your answer by ticking (√) the box.

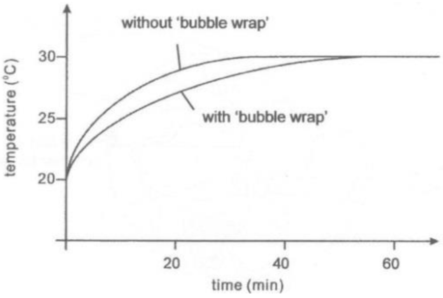
P □ Q □ R □







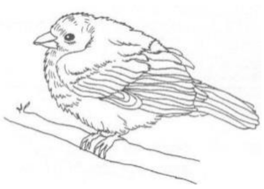
Item 2: John recorded the readings of the thermometer over time. His results are shown below.



Based on his results, John concluded that the beaker with the 'bubble wrap' gains heat more slowly. Is he correct? Circle your answer.

Yes / No / Cannot Tell

Item 3: A bird with a thick layer of feathers is shown below. There are air pockets among the feathers.



Based on the results of John's experiment, explain how the air pockets keep the bird warm in cold air.

Figure 2. Three items of a question in the *Draft PSRT*.

Assessment Structure
(a) Test Booklets

The *Draft PSRT* consisted of five paper-and-pencil test booklets assembled according to multiple-matrix sampling principles (Sirotnik & Wellington, 1977). These broad-ranging principles guided the development and choice of booklet design, such as the breadth of the construct domain to sample, number of reporting scales to include, and the testing duration (Gonzalez & Rutkowski, 2010). Though more commonly deployed in large-scale educational assessments and surveys, this research used a similar design to leverage on key advantages afforded by the deployment of Rasch modelling as the main data analytic method (van der Linden, Veldkamp, & Carlson, 2004). For instance, Rasch modelling facilitated representative testing of important aspects of scientific reasoning with no compromises on testing rigour. Also, there is flexible item assembly and construction as booklets need not each carry the same number of questions or be of the same total marks for psychometric analysis. Moreover, pupils experience less testing fatigue and demotivational effects as they need not attempt all questions to enable inferenece of abilities from the data (Johnson, 1992).

Specifically, the five booklets adopted a balanced and incomplete sampling matrix design (van der Linden et al., 2004). In this design, there is an initial grouping of questions into 11 *question blocks*; each question block consisted of two to three unique questions. Every block appeared an equal number of times in each positition within the five booklets, and each booklet contained a unique subset of the blocks. Other than the anchor block of questions, every test booklet contained four question blocks. Organising the questions thus into blocks facilitated the assignment of questions into various booklets as well as control the representation of tested content across the booklets (Gonzalez & Rutkowski, 2010). Table 2 shows the organisation of question blocks across the five test booklets (Booklets 1 to 5).

Table 2.
Arrangement of anchor block and ten question blocks across five test booklets.

Booklet	Anchor block and question blocks										
1	Anchor	I	II	III	IV						
2	Anchor				IV	V	VI	VII			
3	Anchor	I				V			VIII	IX	
4	Anchor		II				VI		VIII		X
5	Anchor			III				VII		IX	X

Note. I to X are question blocks, each of which is made up of two to three questions.

Of the 11 question blocks in the *Draft PSRT*, one block is in common to every test booklet. This common block, also known as the *anchor block,* helped specify a mutual basis or metrics for comparison with questions not shared in the other booklets. The remaining 10 question blocks rotated among the booklets, with each appearing in two of the five booklets. Linking the different booklets in this manner is necessary for comparison of item difficulties and pupils' abilities using Rasch modelling as the type of items (e.g., item format and response options) and number of total questions varied among booklets (van der Linden et al., 2004). However, the total number of marks differed minimally among the five booklets. This approach ensured that each booklet had a maximum testing duration of an hour. This duration accorded with the usual school practice of administering class assessments of no longer than an hour. Each pupil attempted only one of the five booklets.

(b) Item Format

The *Draft PSRT* contained a total of 79 items in 32 questions. There are more CRSS items to draw on the format's strength of providing stronger inferences about the pupils' reasoning, understanding and application of scientific concepts. See Table 3 for a breakdown of the number of items by item format.

Table 3.

Number of items by item format in the Draft PSRT.

	Item Format		
	Selected-response objective scoring (SROS)	Constructed-response objective scoring (CROS)	Constructed-response subjective scoring (CRSS)
Number of items	15	15	57

(c) Table of Specifications

The knowledge types and scientific practices assessed in the *Draft PSRT* came from a sample of learning oucomes in the primary science curriculum of Singapore (Ministry of Education Singapore, 2013). Similarly, the approximate balance between the knowledge types and scientific practices in the test aligned with current emphasis in the national and school assessments for primary science. For instance, the *Draft PSRT* assigned almost two-thirds of all available marks to the assessment of *content knowledge* assessment. This testing approach for novice learners is in sync with pedgogical principles promoting the initial acquisition of basic understanding of the material world as a "stepping stone" for future learning in science. Correspondingly, the scientific practice of *giving scientific explanations* received almost half of the total marks in recognition of this fundamental skill. The knowledge and scientific practice with the least assessment focus are *epistemic* and *designing and* *evaluating investigation*. A reduced focus on these skills accorded with the teaching emphasis in the curriculum; primary science pupils do not have enough understanding of the content to appreciate the epistemic requirements of science and have little exposure to investigative work. Table 4 presents a table of specifications showing the targeted marks in the *Draft PSRT*.

Table 4.

Table of specifications – mark distribution by knowledge types and scientific practices.

Knowledge Type	% of marks	Scientific Practice	% of marks
Content	60-80	Giving scientific explanation	40-50
Procedural	15-30	Designing and evaluating investigation	15-30
Epistemic	<10	Interpreting and analysing data and evidence	20-40
Total	100%	Total	100%

References

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments, 3*, 125-156.

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29*, 95-110.

Ministry of Education. (2013). *Science Syllabus Primary 2014*. (978-981-05-8232-6). Singapore: Author Retrieved from http://www.moe.gov.sg/education/syllabuses/sciences/files/science-primary-2008.pdf.

Sirotnik, K., & Wellington, R. (1977). Incidence sampling: An integrated theory for "matrix sampling". *Journal of Educational Measurement, 14*, 343-399.

van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement, 28*, 317-331.

Appendix C Two Tables of Statements in *Proforma*

Table 1.

Statements about construct and relationship to specific evidence of content validity.

No.	Statement about Scientific Reasoning Construct	Evidence of Content Validity
1	Scientific reasoning and its sub-scales (knowledge type, practice and sub-practice) represent an important domain of content *of* and *about* science.	Relevance of domain to science
2	Scientific reasoning and its sub-scales specify important reasoning skills and processes that pupils in primary schools should acquire.	Relevance of domain to curriculum
3	Each sub-scale contributes toward the overall notion of scientific reasoning.	Completeness of domain description
4	Scientific reasoning and its sub-scales are clearly written.	Clarity of descriptors about domain
5	The developmental progression outlined in the descriptors of each of the three practices is logical and reasonable.	Appropriateness of domain description
6	Where appropriate, the reasoning skills and processes in the *Draft PSRT* are taught to primary pupils in the curriculum of your country.	Opportunities to learn the domain in school
7	Where appropriate, the reasoning skills and processes in the *Draft PSRT* can be learnt by primary pupils.	Age-appropriateness of domain
8	The distribution of knowledge types and scientific practices assessed in the *Draft PSRT* broadly follow the national assessments of your country.	Appropriateness of the test development process

Table 2.

Statements about items and relationship to specific evidence of content validity.

No.	Statement about Items in *Draft PSRT*	Evidence of Content Validity
1	The concept(s) assessed represent important content *of* and *about* science.	Appropriateness of the test development process
2	The science content in the question is scientifically accurate.	Appropriateness of the test development process
3	The concept(s) assessed are aligned to the curriculum.	Appropriateness of the test development process
4	The sub-scales have been correctly characterised in the question.	Domain representation and domain relevance
5	The level of cognitive demand is identified correctly.	Domain representation and domain relevance
6	Item Stimulus (a) Contains only words that are essential for responding. (b) Vocabulary and sentence structure are grade-appropriate.	Appropriateness of the test development process
7	Item Stem (a) Text is minimal in length, written as plainly as possible. (b) Vocabulary and sentence structure are grade-appropriate. (c) Target content is evident from stem.	Appropriateness of the test development process
8	Visuals (pictures, charts and graphs) (a) Visual(s) used are necessary for responding. (b) Visual(s) clearly depict or provide the intended information and are as simple as possible. (c) Within visuals, contain only text that is essential for responding.	Appropriateness of the test development process

Appendix D The Score Method

Penfield (2003) and Penfield and Miller (2004) developed the score method to produce confidence intervals as an indication of the expected level of error in estimating the population parameter from the mean rating (sample statistic) of a survey question/statement given by 10 to 20 reviewers. This method overcomes statistical limitations such as violation of normality imposed by responses from a small pool of reviewers and skewed response distributions, a situation common in most content validation investigations. Specifically, the response to a rating scale statement, denoted by R, is treated as an interval data rather than ordinal in estimating the population mean, μ_R, using the sample mean, $\overline{R},$ of individual rating scale statements.

The estimation of μ_R using $\overline{R}$, and drawing conclusions about constructs measured by single items is applied widely in many fields such as education and psychology (Miller & Penfield, 2005). For instance, the value of $\overline{R}$ is typically used to represent the difficulty level of a test item along the trait continuum in classical test analyses. Depending on the level of confidence used, confidence intervals estimate the upper and lower limits of the expected distance of $\overline{R}$ to the μ_R. Millar and Penfield (2005) developed a code in SAS, which is a statistical analysis software for calculating the confidence intervals of rating scales in content validation studies. The code was adapted in this study to generate the 95% score (rating value) confidence intervals for the eight survey statements rated by the 18-membered panel about the construct.

References

Miller, J. M., & Penfield, R. D. (2005). Using the score method to construct asymmetric confidence intervals: An SAS program for content validation in scale development. *Behavior Research Methods, 37*(3), 450-452.

Penfield, R. D. (2003). A score method of constructing asymmetric confidence intervals for the mean of a rating scale item. *Psychological Methods, 8*(2), 149-163.

Penfield, R. D., & Giacobbi, P. R. J. (2004). Applying a score confidence interval to Aiken's item content-relevance index. *Measurement in Physical Education and Exercise Science, 8*(4), 213-225.