Title	Development and field-testing of an instrument for rating cognitive demands of mathematical assessment items
Author(s)	Tan Hwee Chiat June, Ng Wee Leng and Paul M. E. Shutler
Source	*The Mathematics Educator, 17*(1), 57-78
Published by	Association of Mathematics Educators, Singapore

The Mathematics Educator
2017, Vol. 17, No.1, 57 - 78

Development and Field-Testing of an Instrument for Rating Cognitive Demands of Mathematical Assessment Items

TAN Hwee Chiat June[1]
National Institute of Education, Nanyang Technological University, Singapore

NG Wee Leng
National Institute of Education, Nanyang Technological University, Singapore

Paul M. E. SHUTLER
National Institute of Education, Nanyang Technological University, Singapore

Abstract: Teachers' judgements of the cognitive demand of mathematical assessment items have implications for the nature of students' learning experiences. However, existing taxonomies for classifying cognitive demand are often not customised for pre-university or A-level mathematics teachers to use. This paper reports on the development and field-testing of a cognitive demand instrument specifically for helping A-level mathematics teachers to sharpen their judgements of the cognitive demand of A-level mathematical assessment items. Fourteen A-level Mathematics teachers from a junior college and an Integrated Programme school in Singapore participated in this study, where they rated the cognitive demand of assessment items on selected Pure Mathematics topics from the national A-level examinations using the cognitive demand instrument. The instrument was found to be useful to the teachers in providing them with an awareness of how the cognitive demand of A-level mathematical assessment items could be understood through the dimensions of complexity, abstractness, and strategy.

Keywords: cognitive demand, mathematics assessment, pre-university mathematics teachers

Introduction

Often after a test or an examination, it is not uncommon for teachers and students to describe a test question or an examination as "easy", "tedious",

[1] This author was a student in the Master of Education (Educational Assessment) programme.

"difficult", or "challenging". However, each of them could in fact be referring to different features of an examination which they consider as 'demands'. Cognitive demand, in particular, refers to the cognitive processes that examiners place on students to perform the tasks set out in the item (Pollitt, Ahmed, & Crisp, 2007). It is inherent in any assessment item and can be manipulated (i.e., by intentionally increasing or reducing the demand level) so that specific thinking processes can be assessed. Thus, cognitive demand plays a significant role in the function and purpose of the assessment.

According to Schoenfeld (1988), teachers rely heavily on the items in standardised tests or national examinations, for example, as a guide on the concepts or skills to emphasise. In the context of Singapore, a common practice among Advanced level (A-level) mathematics teachers is to use items from the General Certificate of Education (GCE) A-level mathematics examinations in their instructional materials. The GCE A-level course is a two-year pre-university course equivalent to Grades 11 and 12. There is no prescribed or standardised set of textbooks for the A-level mathematics curriculum in Singapore, and teachers are responsible for designing their lesson materials and setting school-based assessments. Thus, the GCE A-level examination items also serve as an important point of reference for teachers when they ~~set~~ develop or modify items for their own school tests or examinations. Hence, it is important that mathematics teachers are cognisant of the cognitive demand of the assessment items that they select, construct and use. In this way, they can ensure that items of varying cognitive demand profiles are used to help students develop the cognitive processes needed for A-level mathematics. They can also determine the level of cognitive challenge appropriate for students so as to motivate students in learning and prepare them for the GCE A-level examinations.

Although there are various cognitive demand frameworks or taxonomies such as, Bloom's Taxonomy of educational objectives (Bloom et al., 1956) and the levels of cognitive demand for mathematical instructional tasks by Smith and Stein (1998), the existing frameworks do not provide A-level teachers with a useful tool to help them gauge the cognitive demand of assessment items in A-level mathematics for various reasons. For example, Thompson (2008) found that high school mathematics teachers had difficulty in interpreting and using Bloom's taxonomy to guide their writing of higher-

order thinking questions for an Algebra test. Additionally, they had a tendency to overestimate the level of thinking required in the questions, which could be due to misinterpretations of the cognitive demand categories in the taxonomy.

A problem then emerges that there is currently no tool available that is well suited for helping A-level mathematics teachers to effectively judge the cognitive demand of the assessment items that they select and use to engage students in learning mathematics. As more advanced and abstract topics are covered at A-level, it is essential that students engage in mathematical tasks of varying levels of cognitive demand to support the development of key mathematical processes such as mathematical reasoning and communication. Thus, it is important that A-level mathematics teachers are aware of the demand of assessment items in terms of cognitive processes and not content coverage alone.

This study is part of a broader study on the perspectives of A-level mathematics teachers on the cognitive demand of A-level mathematical assessment items by the authors. To guide the development of a cognitive demand instrument for analysing A-level mathematical assessment items in this study, we posed the following research question: How reliable and appropriate is the instrument for describing the cognitive demand of A-level mathematical assessment items as perceived by A-level mathematics teachers?

With a focus on A-level mathematics teachers, this present study attempts to address the lack of research that investigate teachers' perceptions of the cognitive demand of mathematical items, particularly at A-level. As teachers' awareness and judgements of the cognitive demand of A-level mathematical assessment items have implications for the nature of students' learning experiences and what students come to understand the discipline of mathematics to be, it is worthwhile to develop an instrument for gauging the cognitive demand of A-level mathematical assessment items particularly for A-level mathematics teachers to use. Such an instrument would provide them with a useful tool to help them better evaluate the assessment items by sharpening their judgements of the cognitive demand of the items.

Cognitive Demand Taxonomies

Of the many taxonomies available, the most commonly used is the taxonomy of educational objectives by Bloom et al. (1956). Bloom's Taxonomy, as it is often referred to, provides a means for classifying curricular objectives and test items. It comprises six categories in the cognitive domain: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation, and these are ordered from simple to complex. Bloom's taxonomy was later revised to show its connections with different levels of knowledge (Krathwohl, 2002). As with the original taxonomy, the cognitive process dimensions in the revised taxonomy are considered to be different in terms of cognitive demand, with Remembering at the lowest level and Creating at the highest level.

While Bloom's Taxonomy is not specific to any subject, Smith and Stein (1998) proposed four levels of cognitive demand for classifying mathematical instructional tasks: (1) Memorisation (e.g., reproducing facts, formulae or definitions); (2) Procedures without Connections (e.g., algorithmic, focussed on producing correct answers instead of developing mathematical understanding); (3) Procedures with Connections (e.g., use of procedures for purposes of developing deeper levels of understanding of mathematical concepts and ideas); and (4) Doing Mathematics (e.g., requiring complex and nonalgorithmic thinking). Memorisation and Procedures without Connections are classified as lower-level demand whereas Procedures with Connections and Doing Mathematics are classified as higher-level demand.

Hughes, Pollitt, and Ahmed (1998) offered another perspective on gauging the cognitive demand of assessment items in terms of the complexity of operations, the resources needed, the extent of dealing with abstractness, and the extent to which students need to devise a strategy for solving problems. The CRAS (Complexity-Resources-Abstractness-Strategy) framework, as it is known, was based on the scale of cognitive demand by Edwards and Dall'Alba (1981) and was developed for the purpose of analysing A-level examination items. It has been used, for example, to examine the comparability of qualifications by Crisp and Novaković (2009), but reliability and validity estimates were not reported.

Despite being widely used as a model for identifying cognitive processes, the assumption that there is a hierarchical ordering of categories and that learning outcomes can be classified into one of these categories gives rise to practical difficulties in using Bloom's taxonomy (Kilpatrick, 1993). Considered as low on Bloom's taxonomy, a knowledge-level question, for instance, could be designed as a high cognitive demand question. It could also be that more than one category of the taxonomy could be appropriate for describing the cognitive processes involved in a question with several learning outcomes. Gierl (1997) and Kilpatrick have highlighted other difficulties in using Bloom's taxonomy such as, interpreting the categories of the taxonomy in the context of the subject matter of interest and the independence between content and the cognitive processes. In particular, Kilpatrick noted that a common criticism of Bloom's taxonomy was the amount of time needed by users to relate and adapt the categories of the taxonomy to their subject matter.

While a key criticism of Bloom's taxonomy is that users need to interpret it in the context of the subject that they teach, the taxonomy by Smith and Stein (1998) addresses this issue by being subject-specific, focusing on mathematics. Their taxonomy is oriented around mathematical activities or practices which are grouped under four cognitive levels. This makes it useful for teachers as it gives them an idea of the nature of the mathematical practices that students should ideally engage in, so as to develop the desired cognitive processes. However, this taxonomy is more applicable for examining classroom tasks, where a range of cognitive demands can be built in. It may be difficult to find tasks at the highest level of "Doing Mathematics" in an examination, given the constraints and purposes of an examination or a system-wide assessment.

As with Bloom's taxonomy, the CRAS framework of Hughes et al. (1998) is not subject-specific, and therefore requires users to interpret the dimensions of cognitive demand in the framework in the context of their subject matter. However, the dimensions of cognitive demand considered in the framework are relevant and connected to mathematics learning and assessment. For example, the idea of complexity is evident in the nature of the mathematical procedures that students need to carry out, and the ability to abstract mathematical concepts is an essential skill in learning mathematics. Additionally, for each dimension, a progression of levels of demand is

considered and this helps to support users of the framework in exercising their judgements. Moreover, the CRAS framework was constructed with the purpose of studying the cognitive demand of examination items.

The CRAS framework (Hughes et al., 1998) has useful features that lend itself well to modification or adaptation to obtain an instrument that could be used by teachers for gauging the cognitive demand of assessment items. In this framework, cognitive demand is unpacked into four distinct dimensions. Each dimension is further unpacked and described in terms of a continuum or scale of cognitive demand levels. More importantly, the scales are unidimensional and essentially qualitative in nature. These features not only provide a richer description of the nature of the cognitive demand of an assessment item, they also provide a focus during the decision-making process as users would need to consider only the demand descriptors for one dimension at a time when judging the demand of an item. A richer description of cognitive demand is useful and valuable when analysing assessment items as assessment items range from simple to complex and could have been designed to assess one or more learning objective. With the CRAS framework, problematic issues such as ensuring that items fall only in one category of cognitive demand and that the hierarchical ordering of categories would still hold, as in the case of adapting Bloom's taxonomy, are avoided. Taken together, the CRAS framework seemed the most appropriate for adapting and modifying for teachers to use to analyse the cognitive demand of A-level mathematical assessment items.

Development of the Instrument

To develop the cognitive demand instrument for analysing A-level mathematical assessment items, the CRAS framework (Hughes et al., 1998) was adapted and modified. As previously mentioned, the CRAS was developed to provide examiners with a scale for assessing the levels of cognitive demand placed on students in the context of examinations. In this framework, cognitive demand was conceptualised in terms of four dimensions: complexity, abstractness, resources, and strategy. Within each of these dimensions, there are five levels of demand, of which only two levels (i.e., Level 2 and Level 4) are defined. In total, there are eight descriptors in the CRAS that are intended to facilitate qualitative judgements

by users of the framework on the cognitive demand of assessment items. The eight descriptors served as 'anchor' descriptors, while providing flexibility for users to adapt the framework to the needs of their own subject areas as well as to use their professional knowledge in gauging the cognitive demand of assessment items. Thus, the CRAS is not domain-specific and could be applied to a variety of subjects.

Details of the modifications to the CRAS in developing the cognitive demand instrument (hereafter, referred to as the CD instrument) are described below. A summary of the modifications is shown in Appendix 1, and the CD instrument is provided in Appendix 2.

Modifications to the CRAS Framework

A notable difference between the CRAS and the CD instrument developed for analysing cognitive demand in A-level mathematical assessment items is in the dimensions of cognitive demand considered. Cognitive demand in A-level mathematical assessment items was conceptualised in terms of three dimensions: complexity, abstractness, and strategy. These three dimensions are applicable and relevant to the learning of mathematics. For example, mathematical problem solving involves devising a strategy or an approach to tackle the problem. It also involves a logical sequence of interdependent steps or subgoals that could range from simple to complex. The need to deal with concrete and abstract objects and ideas in solving mathematical tasks, which was based on Novak's (1977) interpretation of Piaget's theory of cognitive development in the CRAS, is also often regarded as a critical aspect of understanding and learning mathematics. Unlike the CRAS framework, the resources dimension was excluded from the CD instrument as students are often not required to generate data or information in A-level mathematics examinations.

The dimensions and descriptors in the CRAS, which can be generalised across a variety of subjects, was specifically interpreted and mapped to the subject of A-level Mathematics. In the CRAS, cognitive demand in the complexity dimension was described in terms of the operations, comprehension needed, and links between the operations. These descriptors were re-interpreted and considered from the perspective of the complexity of cognitive processes placed on A-level mathematics students for the CD

instrument. This led to the definition of complexity in the CD instrument as the demand placed on students in terms of the subgoals or set of steps and the links between them that students were expected to be able to go through when they attempt the mathematical assessment items. Complexity in understanding technical language, as described in the CRAS complexity dimension, was integrated into the abstractness dimension of the CD instrument instead. This was because in the context of mathematics, technical language would likely include the use of mathematical notations and symbols, which in turn, could affect the level of cognitive demand in abstractness instead.

The abstractness dimension in the CRAS was loosely defined in terms of *dealing with concrete objects* (i.e., Level 2) versus *highly abstract* (i.e., Level 4). In the CD instrument, the abstractness dimension was defined more specifically for mathematics, where abstract elements referred to mathematical objects and ideas that required considerable mental construction or some imagination (e.g., unknown functions and limits). On the other hand, concrete elements in mathematics were defined in the CD instrument as mathematical objects that were completely specified (e.g., numerical coefficients). The strategy dimension in the CRAS focussed on the extent to which students have to devise the strategy as well as organise and communicate their answer. In the CD instrument, cognitive demand in the strategy dimension focuses specifically on the demand placed on students in terms of how they decide to tackle the assessment item. Cognitive demand in terms of organisation of solution was not included in the CD instrument as it was generally not heavily or explicitly emphasised in the A-level mathematics examinations (as evident from the specimen paper mark schemes for the various A-level mathematics syllabuses used in Singapore).

As with the CRAS, there are five levels within each dimension in the CD instrument. However, unlike the CRAS, where there are only eight given descriptors, all the five levels within each of the three cognitive demand dimensions in the CD instrument were defined. Each of these levels was defined using comparative language to show a progression with respect to one particular aspect of that demand (i.e., complexity, abstractness, or strategy). For example, the level descriptors in the complexity dimension in the CD instrument ranged from *a single step or very few steps with no linkages between the steps* (Level 1) to *many steps with complex linkages*

between steps (Level 5). For the strategy dimension in the CD instrument, the levels of demand (from 1 to 5) describe the extent to which students are told what to do (either by the question itself or because the procedure is a standard one which they have memorised) or they are left to decide for themselves what to do. The use of comparative language and a focus on one particular aspect of demand helps to ensure unidimensionality and minimise overlaps in meaning across the dimensions.

As the CD instrument was developed based on the CRAS, there are similar features between these two instruments, for example, the scale length, common dimensions of cognitive demand, and the purpose of the instruments (i.e., for judging the cognitive demand of examination items). More importantly, like the CRAS, the CD instrument also assumes that judgements of cognitive demand are not intended to be combined quantitatively to obtain a numeric score of overall demand. According to Johnson and Mehta (2011), combining scores into a single rating of overall demand would result in loss of qualitative information that could be gleaned from considering each dimension separately or by investigating how the dimensions interact with one another.

Field-Testing the Cognitive Demand Instrument

Sample

Fourteen teacher participants from two schools of different profiles in Singapore participated in the field-test or trial of the CD instrument. There were 9 teachers from School A, which is an Integrated Programme school, where students take a six-year curriculum where they bypass the GCE Ordinary level (O-level) examinations and proceed directly to the A-level course. The GCE O-level examinations are national examinations that students in Singapore take at the end of their secondary school education. There were 5 teachers from School B, which is a junior college that offers the two-year A-level curriculum. There was a mix of beginning and experienced teachers in the sample.

Mathematical Assessment Items

Twelve mathematical assessment items were selected from the GCE A-level Mathematics examinations that were held between the years 1999 and 2015. The items cover selected Pure Mathematics topics (e.g., Functions and Graphs, Equations and Inequalities, and Calculus) and a range of cognitive demand profiles. These items were selected as a significant portion of the A-level Mathematics syllabuses was on Pure Mathematics. Both content and construct validity of the items were ensured since the items were from past A-level mathematics examinations for Singapore candidates.

Questionnaire on CD Instrument

As part of the development and field-testing of the CD instrument, teachers' perceptions of cognitive demand were also gathered using a questionnaire. This questionnaire was intended to assess teachers' views on the CD instrument after they had used the instrument to rate the cognitive demand of the assessment items. Their feedback on the instrument could inform future work on refining the instrument. For example, teachers were asked to indicate their extent of agreement to a set of statements pertaining to the usefulness of the instrument, or more specifically, the rubrics (e.g., "The rubrics are useful in helping me to gauge the cognitive demand of questions"; "I am more aware of the cognitive demand of questions after using the rubrics"). Teachers were also asked to rank the three dimensions in terms of the difficulty they experienced in selecting and rating and to explain their reasons.

Procedures and Data Analysis

The teacher participants were informed that the purpose of the field-test was to develop a tool to help A-level mathematics teachers gauge the cognitive demand of A-level mathematical assessment items. A training session (between 50 minutes to one hour) was conducted for the teacher participants from each school on how to apply the CD instrument to rate the cognitive demand of mathematical assessment items. They were briefed on the dimensions of cognitive demand and the level definitions, and had hands-on practice in using the instrument with a set of test items (mainly selected from past A-level mathematics examinations) that were not included in the set of 12 mathematical assessment items selected for the actual field-test.

For example, the teacher participants were given items similar to the one shown in Figure 1.

Sketch the curve with equation $y = \left|\frac{x-3}{x+5}\right|$, stating the equations of the asymptotes. On the same diagram, sketch the line with equation $y = 4 - x$. Hence, solve the inequality $\left|\frac{x-3}{x+5}\right| < 4 - x$.

Figure 1. Example of parallel item used for training.
(Actual item not reproduced due to copyright.)

To rate the level of demand in the complexity dimension, they were asked to consider the number of steps (and linkages between them) that the item-setter required students to go through so as to solve the problem. For the abstractness dimension, participants were asked to consider the extent to which students would need to deal with abstract or concrete elements as intended by the item-setter. For rating the strategy dimension, they had to make a judgement as how much guidance was provided in the item. A possible rating for the item shown in Figure 1 would be: Complexity – Level 3 (i.e. sketch the two graphs and solve the inequality, so there are few steps which are connected); Abstractness – Level 1 (largely concrete mathematical objects/ ideas involved); and Strategy – Level 1 (standard procedure required and very guided).

The teacher participants were also given time to ask questions and clarify their understanding on the use of the CD instrument. At the end of the training session, they were given the set of 12 items for the actual field-test, a form for recording their cognitive demand ratings as well as the questionnaire. The teacher participants were asked to work independently to rate the demand of the items by dimensions using the CD instrument as a guide. They were also informed that they could use their professional judgement in helping them gauge the cognitive demand of the items. They were given about one week to complete their ratings and the questionnaire.

To measure interrater reliability among the teachers, Cohen's kappa (for multiple raters) and Krippendorff's alpha (Krippendorff, 1970) were calculated. Among the advantages of using Krippendorff's alpha are that it measures agreement across any type of data, also it can be used with any number of raters, and furthermore, it uses bootstrapping techniques (Hayes & Krippendorff, 2007). Krippendorff's alpha was computed using the SPSS macro MKappa, while Cohen's kappa for multiple raters was calculated using the macro MKAPPASC.

Pilot Trial of the CD Instrument

Prior to field-testing the CD instrument with the teacher participants, the authors independently rated the set of 12 mathematical assessment items meant for the teachers using the CD instrument to check the clarity of the cognitive demand descriptors. Minor modifications were made to the CD instrument, for example, rephrasing the descriptors for the strategy dimension to improve clarity and to show a clearer progression in terms of the cognitive demand. Interrater reliability among the three authors was high, with Krippendorff's alpha (Krippendorff, 1970) values of .945, .837, and .861 for complexity, abstractness, and strategy, respectively. All 12 items were successfully classified according to the three cognitive demand dimensions. In other words, all the items could be assigned to one of the five levels of demand for each of the three cognitive demand dimensions. Overall, there was a mix of items with high demand in one or two dimensions and lower demand in the remaining dimensions.

Results

Interrater Reliability

All the 14 teacher participants provided their ratings of cognitive demand for complexity, abstractness, and strategy for all 12 assessment items. For overall demand, only 13 of the teacher participants provided ratings as one teacher had submitted a blank form. Although the CD instrument has five levels of demand for each of the three dimensions, in general, the teacher participants tended to use the lower levels of the instrument (i.e., Levels 1 to 3) in their ratings more often compared to the higher levels (i.e., Levels 4

and 5). In fact, there were five teachers who did not assign a Level 5 to any of the items. There was no consensus among the teacher participants in their ratings of cognitive demand for the items. There were items, for instance, where majority of the teachers rated the demand in complexity to be Levels 1 or 2 but a minority judged the item to be at least a Level 3 in demand.

To evaluate interrater reliability, Krippendorff's alpha (Krippendorff, 1970, 2013) and Cohen's kappa for multiple raters were calculated (using the SPSS macros MKappa and MKAPPASC respectively). These estimates of interrater reliability are presented in Table 1. Krippendorff's alpha values ranged from .442 to .547, which were below the recommended cut off of at least .67 or .70 (Krippendorff, 1970, 2013). Considering the less conservative Cohen's kappa as an estimate of interrater reliability, there was slight to fair agreement among the teacher participants as the kappa values ranged from .171 to .292. This was based on the guidelines suggested by Landis and Koch (1977), where kappa values between .00 to .20 indicate slight agreement and values between .21 to .40 indicate fair agreement. Intrarater reliability was not evaluated as the teacher participants were asked to rate the cognitive demand of the assessment items only once, it was not possible to check the intrarater reliability.

Table 1
Interrater Reliability Estimates

Cognitive Demand	Krippendorff's alpha	Cohen's kappa
Complexity	.442 [0.380, 0.504]	.171*
Abstractness	.547 [0.491, 0.601]	.213*
Strategy	.531 [0.473, 0.587]	.245*
Overall	.445 [0.335, 0.549]	.292*

Note. * $p < .001$.

Correlations among Cognitive Demand Dimensions

According to Edwards and Dall'Alba (1981), the interactions of all the dimensions determine the cognitive demand of a task. Thus, Spearman rank-order correlations were also computed to examine the relationships among

the dimensions of cognitive demand and also overall demand. The intercorrelations among these measures of cognitive demand are shown in Table 2. There were significantly strong positive relationships between overall demand and each of the three cognitive demand dimensions. Additionally, complexity, abstractness, and strategy were also highly correlated to one another.

Table 2
Intercorrelations among Cognitive Demand Dimensions and Overall Demand

Cognitive Demand		1	2	3	4
1.	Complexity	—	—	—	—
2.	Abstractness	.753**	—	—	—
3.	Strategy	.858**	.781**	—	—
4.	Overall	.872**	.781**	.846**	—

Note. $^{**}p < .01$.

Usability and Usefulness of the CD Instrument

As the focus of this study was to develop a cognitive demand instrument particularly for the use of A-level mathematics teachers, the views of the teacher participants regarding the usability and usefulness of the CD instrument were gathered. Out of the 14 teacher participants, 12 of them completed the questionnaire on the usefulness and usability of the CD instrument. Overall, very positive feedback was provided by these teachers as shown in Table 3. The teacher participants were also asked if there were any other dimensions of cognitive demand that should be considered in the CD instrument. Seven of them agreed that the three dimensions of cognitive demand were sufficient and no additional dimension needed to be added. However, one teacher participant suggested that students' familiarity with the subject matter and problem solving strategies could be incorporated into the level descriptors, whereas another teacher suggested that there could be tasks with few steps but complex linkages, which do not belong to any of the

complexity demand levels. The remaining three teachers did not provide any response to this question.

Table 3
Summary of Feedback on Usefulness and Usability of CD Instrument (*n* = 12)

Item		SD	D	A	SA
(a)	The rubrics are useful in helping me to gauge the cognitive demand of questions.	0	0	7	5
(b)	I am more aware of the cognitive demand of questions after using the rubrics.	0	0	2	10
(c)	The rubrics cover important dimensions of cognitive demand.	0	0	9	3
(d)	The rubrics can be applied to any topic in the A-level mathematics syllabus.	0	0	9	3
(e)	The rubric descriptors provide a common language for me to discuss the demands of mathematics questions with my colleagues.	0	0	9	3
(f)	I can use the rubrics as a guide when I set questions for tests/exams in future.	0	0	10	2

Note. SD = strong disagree; D = disagree; A = agree; SA = strongly agree.

The teacher participants were also asked to indicate which of the dimensions of cognitive demand they found the most difficult to rate. Abstractness and strategy (five teachers each) were tied as the most difficult dimension to rate. One teacher explained that it was "hard to decide on strategy because there are so many strategies used in mathematics". There were also teachers who commented that ratings for strategy could be influenced by a teacher's own knowledge of problem solving strategies and students' prior exposure to the item. For another teacher, strategy and complexity were interlinked which made the rating process difficult. For abstractness, the teachers explained that this dimension was difficult to rate because the concept of abstractness was vague and there could be more to it than considering "how many unknowns were involved". Seven teachers found the complexity dimension easiest to rate but they did not provide the reasons why they thought so.

Discussion

All items were successfully classified according to the levels of demand for the three dimensions, providing evidence of content validity. In other words, the teacher participants were able to assign one of the levels of demand for each dimension to all the assessment items. However, the teachers differed in terms of the level, from 1 to 5, on which the items were rated. Interrater reliability was below the recommended threshold for establishing consensus agreement or disagreement (based on Krippendorff's alpha), or at best, there was fair agreement beyond chance between pairs of teachers (based on Cohen's kappa). This could be due to the number of levels of demand for each dimension, since it is easier to achieve high reliability with fewer levels in the rubric (Jonsson & Svingby, 2007).

In their development of a cognitive demand scale for secondary science, Edwards and Dall'Alba (1981) found that teachers tended to categorise the type of demand based on how they intended to use the tasks rather than strictly follow the descriptions of cognitive demand in the scale. This could have been the case with the A-level mathematics teachers who participated in this study, even though they were taught how to use the CD instrument. In addition, the teacher participants' knowledge of their students could have interfered with their evaluation of the demand level and they could have rated the items based instead on how successful they thought their students would be in tackling the items. This could also explain why the teacher participants had found the complexity dimension easiest to rate but differed in terms of their ratings which was reflected in the low alpha value. They might also not have been consistent in their judgements of complexity when they analysed the items.

Nonetheless, the results of the interrater reliability suggest a need to further clarify the CD instrument with the teachers so that they will then able to assign items to the level of demand, based on a shared understanding of the cognitive demand dimensions and according to the qualitative descriptions of cognitive demand in the instrument. The high intercorrelations among the dimensions of cognitive demand also point to a need to examine if there is any redundancy in the choice of cognitive demand dimensions. However, as a first trial of such an instrument for A-level mathematics, the reliabilities obtained in the field-testing of the CD instrument are encouraging.

Based on the teacher participants' responses to the questionnaire, the CD instrument proved to be useful to the teacher participants in providing them with an awareness of how the cognitive demand of A-level mathematical assessment items could be understood through the dimensions of complexity, abstractness, and strategy. The dimensions of cognitive demand in the CD instrument seemed to have captured the types of cognitive demand likely to be found in A-level mathematical assessment items, although there were two suggestions on possible aspects of cognitive demand that could have been overlooked. One of the suggestions was to consider students' familiarity with the concepts required to answer the item, but this would not constitute a type of cognitive demand as defined in the instrument. The other suggestion highlighted a possible gap in the complexity dimension in the case of items with few steps but complex linkages. Such items could exist but might not be seen in A-level mathematics examinations. On the other hand, the teacher participant could actually be referring to the complexity of the mathematical operations. However, this could not be ascertained without interviewing the teacher.

The findings from this study have several implications. For teachers, being able to identify the cognitive demand of an item is a useful skill to have in their work and the CD instrument could be a possible tool for teachers to adapt and use in their subject domain. Particularly for A-level mathematics teachers, the CD instrument provides them with a common language and reference scale to describe and discuss the cognitive demand of assessment items in mathematics. With the instrument, they can analyse and discuss items in terms of the cognitive processes that students should develop or probable causes of cognitive difficulty. They can also be more aware of the cognitive processes required in solving mathematical assessment items and sharpen their own judgements of what constitutes demand in A-level mathematics. In this way, teachers can develop their knowledge and expertise in teaching and improve their assessment practices by assessing the full range of performance of students.

This study also offers curriculum planning or assessment officials in education ministries a glimpse of how teachers' perceptions of the nature of the cognitive demand of the A-level mathematical assessment items could be gathered with the CD instrument, which was specifically developed for teachers. Teachers' perceptions or ratings of cognitive demand could be

used by those involved in curriculum and assessment work as a springboard for discussions on the effectiveness of the curriculum being implemented and the quality of assessment practices in schools.

Additionally, there are practical implications for researchers and examination developers. This study builds on the work of Hughes et al. (1998) by defining what constitutes cognitive demand for A-level mathematics and in developing a tool for gauging cognitive demand for teachers. Studies on cognitive demand usually involve examiners or subject matter experts. Having a tool that could be used by teachers could offer new perspectives on cognitive demand as understood by teachers, who have the most interactions with students in schools. Finally, researchers and examination developers may also need to consider if the CD instrument has the value to be another method or tool that is potentially useful for judging the cognitive demand of assessment items.

Conclusion

Selecting and developing assessment items is an integral part of a teacher's work. Thus, the CD instrument developed in this study has the potential to be useful for A-level mathematics teachers. More specifically, based on the feedback from the teacher participants, the CD instrument provided a starting point in enabling teachers to gauge the cognitive demand of mathematical assessment items in a more focussed way, giving greater clarity and precision to their judgements. In addition, the use of the CD instrument provided insights into the dimensions that constitute cognitive demand, allowing comparisons of cognitive demand to be made across A-level mathematical assessment items on different topics and different profiles of demands.

Although a small sample size was used in this study, the CD instrument showed potential for teachers to use it to guide their selection and design of items on Pure Mathematics topics as well as to strengthen their teaching of related concepts and processes. Future research could be done to examine how the validity and reliability of the CD instrument can be improved. For example, further studies can be conducted to investigate how the CD instrument can be used by teachers to achieve greater reliability in their

ratings of cognitive demand and to establish a common understanding of the cognitive demand in A-level mathematical assessment items. A larger sample could be used and a more representative sample of teachers from the population of A-level mathematics teachers across different schools could be considered. Since the CD instrument is intended to capture the cognitive demand in the A-level mathematics curriculum, the set of topics or items could also be broadened. Finally, the potential use of the CD instrument to primary and secondary mathematics teachers could also be examined in future research.

With a greater awareness of cognitive demand, both individually and among colleagues at the departmental level, teachers can be more confident not only in setting test and examination items but also in determining how well the assessment that they have designed is aligned to the curriculum. More importantly, their judgements of cognitive demand can also be used to guide the teaching of important mathematical processes. Overall, a cognitive demand instrument customised for A-level mathematics teachers could help them to be more conscious of the dimensions of cognitive demand inherent in a range of mathematical tasks or assessment items and, in particular, in sharpening their judgements of the demand of assessment items.

References

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York: Longmans, Green.

Crisp, V., & Novaković, N. (2009). Is this year's exam as demanding as last year's? Using a pilot method to evaluate the consistency of examination demands over time. *Evaluation and Research in Education*, *22*(1), 3–15.

Edwards, J., & Dall'Alba, G. (1981). Development of a scale of cognitive demand for analysis of printed secondary science materials. *Research in Science Education*, *11*(1), 158–170.

Gierl, M. J. (1997). Comparing cognitive representations of test developers and students on a mathematics test with Bloom's Taxonomy. *The Journal of Educational Research*, *91*(1), 26–32.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89.

Hughes, S., Pollitt, A., & Ahmed, A. (1998, August). *The development of a tool for gauging the demands of GCSE and A level exam questions*. Paper presented at the British Educational Research Association Annual Conference, Queen's University Belfast, UK.

Johnson, M., & Mehta, S. (2011, June). Evaluating the CRAS framework: Development and recommendations. *Research Matters: A Cambridge Assessment Publication*, *12*, 27–33. Retrieved from http://www.cambridgeassessment.org.uk/Images/ 109988-research-matters-12-june-2011.pdf

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144.

Kilpatrick, J. (1993). The chain and the arrow: From the history of mathematics assessment. In M. Niss (Ed.), *Investigations into assessment in mathematics education: An ICMI study* (pp. 31–46). Dordrecht: Kluwer Academic Publishers.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, *41*(4), 212–218.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, *30*(1), 61–70.

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology*. Los Angeles; London: SAGE.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Novak, J. D. (1977). *A theory of education*. London: Cornell University Press.

Pollitt, A., Ahmed, A., & Crisp, V. (2007). The demands of examination syllabuses and question papers. In P. Newton, J. A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 166–211). Retrieved from https://www.gov.uk/government/ publications/techniques-for-monitoring-the-comparability-of-examination-standards

Schoenfeld, A. H. (1988). When good teaching leads to bad results: The disasters of "well-taught" mathematics courses. *Educational Psychologist*, *23*(2), 145–166.

Smith, M. S., & Stein, M. K. (1998). Selecting and creating mathematical tasks: From research to practice. *Mathematics Teaching in the Middle School*, *3*(5), 344–350.

Thompson, T. (2008). Mathematics teachers' interpretation of higher-order thinking in Bloom's Taxonomy. *International Electronic Journal of Mathematics Education*, *3*(2), 96–109.

Authors:

Tan Hwee Chiat June (corresponding author: june.thc@gmail.com**)**, National Institute of Education, Nanyang Technological University, Singapore. **Ng Wee Leng** (weeleng.ng@nie.edu.sg), National Institute of Education, Nanyang Technological University, Singapore. **Paul M. E. Shutler** (paul.shutler@nie.edu.sg), National Institute of Education, Nanyang Technological University, Singapore.

Appendix 1

Features	**CRAS Framework**	**CD Instrument**
Dimensions of Cognitive Demand	4 dimensions: • Complexity • Resources • Abstractness • Strategy Definitions of dimensions not domain-specific.	3 dimensions: • Complexity • Abstractness • Strategy Defined differently and specifically for A-level mathematics. Resource dimension excluded as limited opportunity for assessing such demand.
Levels and Descriptors	5 levels of demand 2 defined descriptors for each dimension	5 levels of demand 5 descriptors for each dimension
Context of Use	Intended for analysing cognitive demand of assessment items in any subject	Specifically for analysing assessment items in A-level mathematics
Assumptions	Descriptors provide a qualitative account of cognitive demand. Relies on users being able to relate their subject-specialist knowledge to the underlying features of the dimension scales. Scores or ratings are not intended to be combined quantitatively to obtain a numeric score of cognitive demand for an entire paper.	

Appendix 2

Dimension	1	2	3	4	5
COMPLEXITY Refers to the demand placed on students in terms of the subgoals or set of steps, specified (or implied) in a question to guide them. The levels 1 to 5 describe the extent to which these steps are very few or many in number, and the extent to which these steps are independent or interconnected.	a single step or very few steps, with no linkages between steps	a few steps, with no linkages between steps	a few steps, with simple linkages between steps	several steps, with fairly simple linkages between steps	many steps, with complex linkages between steps
ABSTRACTNESS Refers to the demand placed on students in terms of working with abstract elements, which are mathematical objects and ideas that require considerable mental construction or some imagination (e.g., variable coefficients, unknown functions, asymptotes and limits), rather than concrete elements, which are completely specified (e.g., numerical coefficients, visible intersections). The levels 1 to 5 describe the relative proportions of abstract versus concrete elements in a question.	All, or almost all, concrete elements	Mostly concrete, and a few abstract elements	Some abstract, and some concrete elements	Mostly abstract, and a few concrete elements	All, or almost all, abstract elements
STRATEGY Refers to the demand placed on students in terms of how they decide to tackle a question. The levels 1 to 5 describe the extent to which the students are told what to do (either by the question itself or because the procedure is a standard one which they have memorised) or they are left to decide for themselves what to do.	strategy completely specified, with step by step instructions given, or involving standard procedure which students have memorised	strategy mostly specified, with much guidance given on how to proceed, leaving very few choices to be made as regards strategy	strategy somewhat specified, with some guidance given on how to proceed, leaving a few choices to be made as regards strategy	strategy mostly unspecified, with little guidance given on how to proceed, leaving some choices to be made as regards strategy	strategy completely unspecified, with no guidance given on how to proceed, leaving many choices to be made as regards strategy