Analysing The Cognitive Demand Of Reading, Writing And Listening Tests

Jackie Greatorex, Cambridge Assessment, United Kingdom
Vikas Dhawan, Cambridge Assessment, United Kingdom

ABSTRACT

The aim of the research was to ascertain whether the questions from particular reading, writing and listening tests elicited a spread of cognitive demand types and whether the questions from the question papers that offered higher grades were of a greater cognitive demand. The research was undertaken in the context of language qualifications taken by candidates from around the world. Six senior examiners used a five point scale to rate the cognitive demands of each question on each of the five dimensions – cognitive complexity, the resources the candidates use, the level of abstractness, the cognitive strategies used to understand the task and how to construct a response. The study was designed to return thirty ratings per question. The ratings were analysed using descriptive statistics. Generally, each component elicited a spread of cognitive demands and the questions from question papers offering higher grades were associated with greater cognitive demand. These findings are in line with good principles of test design. The questions in the listening question papers were less demanding than the questions in the reading and writing question papers, for all dimensions except cognitive complexity. Therefore more demanding listening texts were introduced.

Literature suggests that it is good practice for participants to develop a shared understanding of the cognitive demands for the target assessment before they individually rate questions. This shared understanding may be represented in supplementary material such as key examples of higher and lower demands from the target assessment. However, few studies detail how to develop a shared understanding and the supplementary material is infrequently reported. Therefore the present research gives a method for participants developing a shared understanding of the cognitive demands and reports an illustration of the supplementary material.

Introduction

The aim of the research is to ascertain whether the questions from particular reading, writing and listening question papers elicited a spread of cognitive demand types and whether the questions from question papers which offered higher grades elicited greater cognitive demand than other questions. The study provides a backdrop for:

- Imparting a method of participants developing a shared understanding of cognitive demands which they then used to rate the cognitive demands of questions

- Reporting the supplementary material which partly represents the participants' shared understanding.
- It is important to report the method and supplementary material as few studies report these details.

A major theme in this research is comparing the cognitive demand of questions from different tiers. Tiering is a way of achieving differentiated assessment within an examination system. Teachers allocate candidates to one of a small number of tiers. Each tier consists of several question papers giving access to a limited number of grades. Tiering is a means of offering summative assessments to a wide ability range. It is a widespread practice, for example, (1) National Curriculum tests (taken by 14 year olds in England in English, Mathematics and Science between 1998-2008) and (2) General Certificate of School Examinations (public examinations taken by most sixteen year olds in England). Candidates can only be entered in one tier. In the UK the regulator of school level-qualifications (Ofqual) believes that tiering should only be used when absolutely necessary (Long, 2016).

The benefits of tiering are that candidates do not spend time answering questions that are far too (un)challenging for them and the quality of measurement can be enhanced (Oates, 2013). England experienced some problems with tiering.

First, a small minority of lower tier candidates had their achievement capped at grade C when their work was the standard of grade B on the higher tier (Wheadon & Béguin, 2010). Secondly, Good and Cresswell (1988) found that a higher proportion of candidates achieved each grade boundary (cut score, the lowest mark needed to gain a particular grade) on lower tiers than higher tiers because of judgements when the mark scale was divided into a mark range for each grade. Subsequently, the procedures were changed to guard against such an occurrence (also known as the Good and Cresswell effect). Thirdly, there was some evidence that teachers underestimated black students' final achievement and conflated girls' confidence with ability, consequently these groups were allocated to lower tiers than was appropriate (Elwood & Murphy, 2001).

Good test design ensures that the grade boundaries shared by tiers are comparable and that higher tier questions are of greater demand than lower tier questions. One approach is to have common (overlap) questions in the higher and lower tiers, then consider performance on common questions to help set grade boundaries (Burghes, Roddick, & Tapson, 2001; He, Opposs, Glanville, & Lampreia-Carvalho, 2015). Additionally, setters can use their teaching and examining experience to set questions which are appropriate for higher and lower ability candidates.

Cognitive demand

Cognitive demand is the level of knowledge and skill needed to answer a question successfully (Cambridge Assessment, 2010). Several researchers used a variety of frameworks of cognitive demand to classify assessment tasks.

Shute (1979) designed a framework for classifying the cognitive demands of the questions in an accounting examination. Levels of cognitive demand were distinguished primarily in terms of the volume of ideas and connections between ideas as well as the level of abstractness encountered by examinees completing a question. The work was based on Piaget's theory of cognitive development. Shute's framework was used by Baker and Simon (1985) to compare the cognitive demand of questions from two accounting examinations.

Domyancich (2014) aimed to help academics (teaching first year chemistry undergraduates) to adapt their original multiple choice questions into more demanding questions. He evaluated both the original and revised questions using Zoller et al's framework of cognitive demand to show the difference in demand. The framework focused on the links between concepts and the extent to which the student used a prescribed routine or formed their own strategy for solving the assessment task (Zoller, 1995).

Hale et al. (1995) studied the writing tasks in assessments from 162 degree courses including a variety of subjects such as business, chemistry, civil engineering, computer science, psychology and economics. The writing tasks were used to develop a classification system. Writing tasks were classified in several ways, such as whether they were undertaken as a classroom exercise or a final examination and length of the writing e.g. number of words/pages. Additionally the writing tasks were classified in terms of cognitive demands, which emphasised the thinking skills used to complete the assessment. This conceptualisation of cognitive demand drew heavily from Bloom's taxonomy.

The Mathematics Task Analysis Framework allowed researchers to analyse the cognitive demand levels of tasks in several situations, in the instructional materials, as set up by the teacher and as enacted by students (Stein, 1996). The different levels of cognitive demand were principally characterised by the number of concepts used, the amount of linkage between concepts and the extent to which the task indicates what the student needs to do. The framework built on Bloom's taxonomy.

Jones and Tarr (2007) used the Mathematics Task Analysis Framework to compare the cognitive demands of mathematics tasks in mathematics text books over a period of time. Yang (2009) used the framework to analyse the cognitive demand of tasks as set by the teacher and then as enacted by the students. The research focused on five mathematics tasks set by the teacher and enacted by junior high school pupils in China. In two cases the cognitive demand declined between the task being set and enacted, and in three cases the high cognitive demand designed into the task was maintained when the students attempted the tasks.

Tekkumru-Kisa, Stein, and Schunn (2015) successfully adapted the Mathematics Task Analysis Framework for use in science. The Science Task Analysis framework had two dimensions, one was cognitive demand and the second dimension was scientific knowledge and practice. The levels of cognitive demand were defined in a similar way to the levels in the Mathematics Task Analysis Framework.

Maab (2010) developed a way of classifying mathematical modelling tasks, one aspect of the classification framework is level of cognitive demand. The cognitive demand levels were frequently distinguished in terms of the number of mental objects needed and how the student is required to deal with resources such as tables. The classification system was built from literature. A programme was devised to teach mathematical modelling to 12 year olds (low achievers) and the classification system was used to design the tasks.

Manwaring (2006) explored how pupils work at different levels of cognitive demand, when responding to the same task. Her study was set in the context of the negotiation curriculum for ten and eleven year olds in the USA. Manwaring (2006) viewed cognitive demands as the level and type of thinking required by a task. Levels were largely differentiated by the level of abstraction involved. These levels apply to the cognitive, inter-personal and intra-personal domains. The levels were based on Piaget’s developmental stages and epistemological frames illuminating how people think and feel (Kegan, 1982).

Zohar, Schwartzer, and Tamir (1998) compared the cognitive demand of the questions used in homework, classroom discourse, homework assignments and tests, and at different levels of schooling. The context was biology in junior and high schools in Isreal. Zohar et al (1998) devised a way of classifying questions according to the cognitive demands they placed on students. The cognitive demand levels were characterised by the number of concepts and associations between concepts involved in the question, level of abstraction and whether the student must produce a unique communication such as a speech. The underpinning theory was Bloom’s taxonomy.

Hughes, Pollitt, and Ahmed (1998) devised a scale of cognitive demands called CRAS – Complexity, Resources, Abstractness and (Task and Response) Strategy, which was subsequently updated (Pollitt, Ahmed, & Crisp, 2007), and summarised elsewhere (Crisp & Novaković, 2009a, 2009b). These dimensions, described below, resonate with the conceptualisations of cognitive demand in the aforementioned studies.

- **COMPLEXITY - the number of elements and associations between elements that the candidate must hold in mind whilst undertaking the assessment task. Low demand is when the candidate only has to hold in mind few elements and associations between elements to correctly complete the assessment task. High demand is when the candidate needs to consider many elements and interconnections between elements in order to successfully finish the assessment task.** Complexity is a feature of cognitive demand (Baker & Simon, 1985; Domyancich, 2014; Hale et al., 1995; Jones & Tarr, 2007; Maab, 2010; Zohar et al., 1998).

- **RESOURCES - the proportion of material the candidate needs to finish the assessment task which is given by the question rather than generated by the candidate. Low demand is when the candidate gains most or all information for correctly finishing the assessment task from the question, and the candidate has little need to choose the information. High demand is when the candidate needs to produce most or all information to successfully finish the assessment task, or must choose appropriately from the information provided.** Resources are a property of cognitive demand (Hale et al., 1995; Maab, 2010).

- **ABSTRACTNESS - the degree to which an assessment task entails working with concepts as opposed to concrete articles. Low demand is when the candidate exclusively works with concrete articles to correctly complete the assessment task. High demand is when the candidate purely works with abstract information to successfully finish the assessment task.** Abstractness is a characteristic of cognitive demand (Baker & Simon, 1985; Dunworth, 2008; Hale et al., 1995; Yang, 2009; Zohar et al., 1998). A theoretical underpinning for Abstractness is that higher stages of child and adult development involve dealing with abstract phenomena and lower stages with concrete articles (Kegan, 1982; Piaget, 1954).

- **TASK STRATEGY - the degree to which the candidate must form or choose and sustain their own strategy for solving the assessment task. Low demand is when the candidate can follow the strategy**

provided by the assessment task, there is no need for the candidate to check their strategy and they choose minimal material to correctly finish the assessment task. High demand is when the candidate must form their own strategy and check their use of the strategy to successfully complete the assessment task. Task Strategy is a dimension of cognitive demand (Domyancich, 2014; Jones & Tarr, 2007; Tekkumru-Kisa et al., 2015; Yang, 2009).

- **RESPONSE STRATEGY - the degree to which the candidate must form their own strategy for conveying their response. Low demand is when the candidate has little need to arrange their response to correctly complete the assessment task. High demand is when the candidate needs to arrange their response to successfully finish the assessment task.** Response Strategy is an attribute of cognitive demand (Jones & Tarr, 2007; Tekkumru-Kisa et al., 2015; Zohar et al., 1998).

The dimensions in the CRAS scale each have a different focus but they are not mutually exclusive. Each dimension is a continuum from high to low demand, and has five categories or a scale of 1 to 5 (1=low demand; 5=high demand). Participants rate the cognitive demand of a question (assessment) by assigning it a category from each dimension. Once the ratings are analysed they can be used to identify and compare cognitive demands in different assessment tasks, lesson content and so on (Hughes et al., 1998). An alternative to rating is that participants undertake paired comparisons, meaning that they are presented with a pair of questions (assessments) and judge which is more demanding (Crisp & Novaković, 2009b). The process is repeated for many pairs. Johnson and Mehta (2011) explain that aggregating data from different dimensions does not give the overall demand of an assessment task or examination paper.

Much research concerning the cognitive demands of school level qualifications uses CRAS in a variety of subjects (Crisp & Novaković, 2009a, 2009b; Greatorex, Shaw, Hodson, & Ireland, 2013; Johnson & Mehta, 2011; Pollitt et al., 2007; Shaw & Crisp, 2012).The CRAS framework should be tailored to the target subject (Hughes et al., 1998; Johnson & Mehta, 2011; Pollitt et al., 2007), for example Hughes et al. (1998) devised and reported a version of CRAS for geography, history and chemistry. Furthermore, it is advised that participants agree a shared understanding of CRAS for the target assessments before applying the scale (Crisp & Novaković, 2009a, 2009b; Greatorex et al., 2013). If this advice is followed then there is a unique shared understanding of CRAS for each research project, which may be captured as supplementary material. We found two studies, namely Crisp and Novaković (2009a, 2009b), which developed a shared understanding of CRAS to use with paired comparisons. We found no published articles where participants developed a shared understanding of CRAS before *rating* and reported the supplementary material. Therefore, a focus of the present paper is detailing a method of developing a shared understanding of cognitive demands and including the supplementary material.

Context

The research was conducted in the context of two language qualifications for 14 to 16 year olds from around the world. The qualifications were developed and administered by Cambridge International Examinations, a division of Cambridge Assessment and a department of the University of Cambridge.

The summative assessments in reading, writing and listening were marked by external examiners. The examinations were tiered. Component 1 and Component 2 were reading and writing question papers. Components 3 and 4 were listening question papers. The lower tier comprised components 1 and 3 and spanned grades C to G. The higher tier contained components 2 and 4 and covered grades A* to E. A* was the highest grade available. Each component constituted a question paper. Component 1 shared several common questions with component 2, similarly component 3 and component 4 had common questions. Each component had time zone question paper variants. 'Time zone' question papers were variants of the same question paper taken by candidates who resided in different time zone bands.

In addition to the reading, writing and listening question papers the qualifications included an oral assessment. In the case of one qualification each successful candidate was awarded one grade derived from their marks on all assessments. In the case of the second qualification each successful candidate was awarded two grades, one grade for the oral assessment and a second grade derived from their marks on the reading, writing and listening assessments.

Given this arrangement the oral assessment was beyond the scope of this study.

The research focused on the summer 2014 examination session. This research was undertaken in addition to the awarding body's operational procedures and formed part of a wider validation programme (Shaw & Crisp, 2012). Within this context the following research questions were addressed:

1. Do the questions in each question paper elicit a spread of cognitive demand?
2. Do the question papers within each tier elicit similar levels of cognitive demand?
3. Do the higher tier question papers elicit higher cognitive demands than the lower tier question papers?

The research strategy involved three stages. In stage 1 the researchers decided whether the CRAS was conceptually related to the construct examined by the target assessments, as advised by Johnson and Mehta (2011). In stage 2 participants developed a shared understanding of CRAS for the target assessments, the method draws from Crisp and Novaković (2009a, 2009b). In stage 3 participants rated the cognitive demands of each question.

Method

Participants

Six examiners were recruited. These examiners were recommended by the awarding body for their relevant experience in marking and setting questions.

Procedure

Stage 1 Mapping CRAS to the target assessment

The research team decided that CRAS was suitable for use in the research as the five dimensions were conceptually related to the construct assessed by the question papers.

Stage 2 Developing a shared understanding of CRAS for the target assessments

Each participant received:

- instructions and recording sheets
- the framework of cognitive demands (Pollitt et al., 2007)
- four pairs of questions each from the June 2014 question papers (time zone variant 2)
- a unique set of questions, with two pairs from reading and writing question papers and two pairs from listening question papers.
-
- Participants were asked to note at least three similarities and at least three differences in demands between the two questions in each pair. Next, they used the original CRAS description (Pollitt et al., 2007) to categorise what they had written as:
- low or high demand
- Complexity, Resources, Abstractness, Task Strategy or Response Strategy.
- If the data did not fit one of the existing dimensions then it was recorded as an additional dimension. Participants returned the data to the research team.

The authors and an additional researcher with expertise in CRAS read the qualitative data. The researchers checked whether the data was categorised appropriately and decided whether the data was already encapsulated in the CRAS descriptions given in Pollitt et al. (2007). If the data added to the CRAS descriptions then the data was summarised as supplementary material (Table 1).

Table 1 Supplementary material

	Low Demand 1	**2**	**3**	**4**	**High Demand 5**
Complexity	←		↔	More/lengthier steps needed to complete tasks. E.g. receptive reading/listening.	→
Resources	←	Students do not need to add their own ideas. Data/information is often to be found in a short space of text*. Stimulus material uses a simple language structure. Student is guided to relevant information in the stimulus.	↔	Students need to search within a larger text* area to identify the relevant information. Dismisses information that is irrelevant to answering the question. Stimulus material uses a complex language structure.	→
Abstractness	←		↔	E.g. ideas or reasons.	→
Task Strategy	←		↔		→
Response Strategy	←	E.g. numbered response lines.	↔	Complexity of language used by the student.	→

*written or auditory

Participants attended a one day workshop. The purpose of the workshop was to further develop a shared understanding of Complexity, Resources, Abstractness, Task Strategy and Response Strategy in the context of the target examination. At the workshop the participants were briefed on the CRAS descriptions and the supplementary material. They were provided with the CRAS descriptors and the supplementary material, as well as questions from the June 2014 question papers (time zone variant 2). The participants undertook two exercises. In the first exercise they were asked to individually rate five questions on each of the five dimensions. Then they discussed why they had given particular ratings to each question. In the second exercise the participants worked in small groups. Within each group they discussed ratings for a further five questions until an agreement was reached. The decisions were displayed, for all participants to see. The small groups each explained how they had reached these particular ratings.

Stage 3 CRAS ratings

After the workshop the participants received:

- instructions and recording sheets
- the framework of cognitive demands (Pollitt et al., 2007)
- supplementary material
- June 2014 question papers (time zone variant 1)

Participants individually rated each question on a scale of 1 to 5 (1=low demand; 5=high demand) on each CRAS dimension. They were asked to rate the questions using the original CRAS descriptions together with the supplementary material and their experience of the workshop. The research was designed to gain thirty ratings per question.

Analysis of CRAS ratings

The frequency and percentage of ratings for each dimension and component was calculated. Within each tier the distribution from the reading and writing question paper was compared with the distribution from the listening question paper. The two question papers within a tier were expected to gain a broadly similar percentage of ratings 4 and 5 for each dimension. Within each domain (reading and writing/ listening) the distribution from the higher tier question paper was compared with the distribution from the lower tier. For each dimension the higher tier question paper was expected to gain a greater percentage of ratings 4 and 5 than the lower tier question paper.

Results

Table 2 gives the percentage of ratings for the questions in each component.

Table 2 Percentage of ratings

Component	CRAS dimension	Lower demand			Higher demand		Total
		1	2	3	4	5	
1 Reading and Writing Lower Tier	Complexity	9.85	31.82	36.36	16.67	5.30	100
	Resources	9.09	27.27	39.39	21.97	2.27	100
	Abstractness	16.67	35.61	24.24	16.67	6.82	100
	Task Strategy	14.39	25.00	43.18	10.61	6.82	100
	Response Strategy	14.39	62.12	6.06	6.82	10.61	100
2 Reading and Writing Higher Tier	Complexity	10.00	31.33	36.00	16.00	6.67	100
	Resources	8.67	28.00	33.33	24.67	5.33	100
	Abstractness	17.33	32.67	28.00	14.00	8.00	100
	Task Strategy	14.67	23.33	42.67	10.00	9.33	100
	Response Strategy	16.00	62.00	4.67	7.33	10.00	100
3 Listening Lower Tier	Complexity	12.50	14.58	47.92	25.00	0.00	100
	Resources	5.56	20.83	50.00	23.61	0.00	100
	Abstractness	36.81	32.64	12.50	18.06	0.00	100
	Task Strategy	10.42	20.83	68.06	0.69	0.00	100
	Response Strategy	57.64	38.89	3.47	0.00	0.00	100
4 Listening Higher Tier	Complexity	1.75	18.42	42.98	35.96	0.88	100
	Resources	2.63	31.58	35.96	29.82	0.00	100
	Abstractness	15.79	42.98	21.05	18.42	1.75	100
	Task Strategy	2.63	21.05	59.65	15.79	0.88	100
	Response Strategy	7.89	71.93	10.53	9.65	0.00	100

All components elicited cognitive demands from all CRAS five dimensions, with a spread of higher ratings (4 and 5) and lower ratings (1 and 2) (Table 2). The only exception was component 3 which did not elicit any higher cognitive demands for the Response Strategy (Table 2). Therefore, the question papers generally elicited a spread of cognitive demands.

In each tier the reading and writing question paper elicited a greater percentage of higher cognitive demand ratings than the listening question paper for Resources, Abstractness, Task Strategy and Response Strategy (Table 2). In the

case of Complexity each listening question paper elicited a greater percentage of higher cognitive demand ratings than the corresponding reading and writing question paper. The cognitive demand associated with the reading and writing question papers was considered to be more suitable than that of the listening question papers. Based on our results action was taken to increase the demand of the listening question papers by introducing more demanding resources.
For reading and writing the higher tier question paper elicited a greater percentage of higher cognitive demand ratings than the lower tier question paper for Complexity, Resources and Task Strategy. However, the result was reversed for the Abstractness and Response Strategy. For listening the higher tier question paper elicited a greater percentage of higher cognitive demand ratings than the lower tier question paper for all dimensions. The greater cognitive demand generally associated with the higher tier was in line with good principles of test design.

Discussion

The research explored whether the questions from reading, writing and listening question papers elicited a spread of cognitive demand types, whether the higher tier questions were of a greater cognitive demand, and whether question papers in the same tier were of comparable cognitive demand. These question papers were part of language qualifications taken by candidates from around the world. Participants rated the cognitive demands of each question on each of the five dimensions – Complexity, Resources, Abstractness, Task Strategy and Response Strategy. A purpose of the paper was to describe how participants developed a shared understanding of CRAS, and provide an illustration of supplementary material which represents that shared understanding.

When considering the quantitative findings it is important to note that the categories in a CRAS scale are ordinal and the size of the difference between categories can vary. Therefore, the findings indicate relative demand, rather than a standardised measure of the size of demand. Our research shows which questions are more (or less) demanding. Despite this caveat our findings may prove useful in validating (or investigating) test design/development. A key research finding is that the higher tier questions elicited greater cognitive demand than the lower tier questions, generally. Analysis also shows that generally each component elicited a spread of cognitive demands. These findings accord with good principles of assessment development.

The reading and writing component was judged to be more demanding than the listening component in each tier, for all CRAS dimensions except Complexity. As a result of our findings action was taken to increase the demand of the listening question papers by introducing more complex texts for candidates to listen to.

A focus of the paper is to describe how a shared understanding of CRAS was developed. The method for developing the shared understanding drew from Crisp and Novaković (2009a, 2009b). The main commonalities were the tasks prior to the workshop. However, the methods also diverged in several ways. First, our participants rated the demands, whereas Crisp and Novaković's (2009a, 2009b) participants compared pairs of assessments and decided which was the more demanding on each dimension. Secondly, our participants each judged individual questions, whereas Crisp and Novaković's (2009a, 2009b) participants judged the assessment materials for a whole unit. Thirdly, at our workshops the participants rated both individually, and as a group, whereas Crisp and Novaković's (2009a, 2009b) participants worked individually. Both our and Crisp and Novaković's (2009a, 2009b) approaches appeared to develop a shared understanding of CRAS for the target assessments and build participants' confidence for rating or making paired comparisons. This suggests that the pre-workshop activities are key to developing a shared understanding.

Community of practice literature draws from Lave and Wenger (1991). A community of practice is a network of people with an interest in a particular domain, who participate in a communal activity and continuously negotiate community knowledge and practice. Such communities may be intentionally orchestrated for the purpose of sharing knowledge. This description of learning in a community of practice resonates with our workshop method. The participants were intentionally brought together to participate in the communal activities of:

- discussing the CRAS descriptors, supplementary material and examination questions
- rating questions using the CRAS descriptors and supplementary material

The intention was that the participants would negotiate how to understand the CRAS descriptors and supplementary material. We believe that the participants learnt from each other and negotiated a shared understanding.

Another goal of the paper was to provide an illustration of supplementary material representing the shared understanding of CRAS. The supplementary material from the present research is given in Table 1. The supplementary material is unique to this research and the target assessments.

Practice involves both tacit knowledge that is instinctive and commonly held within the community of practice, as well as knowledge which is explicit (Wenger, 1998). Therefore, not everything can be expressed in text. Consequently, the supplementary material comes with the caveat that some shared understanding of the CRAS dimensions will remain tacit amongst the participants.

In conclusion, when using CRAS we recommend:

1. developing a shared understanding of CRAS for the target assessments
2. participants discuss the CRAS dimensions, supplementary material and communally practice using them before individually rating questions (assessments)
3. reporting the supplementary material.

References

Baker, R. E., & Simon, J. R. (1985). An assessment of the cognitive demands of the uniform CPA examination and implications for CPA review/preparation courses *Journal of Accounting Education, 3*(2), 15-29. doi: 10.1016/0748-5751(85)90003-X

Burghes, D., Roddick, M., & Tapson, F. (2001). Tiering at GCSE: is there a fairer system? *Educational research, 43*(2), 175-187. doi: 10.1080/00131880110051164

Cambridge Assessment. (2010). Cambridge Assessment Exam standards: the big debate. Report and Recommendations. http://www.cambridgeassessment.org.uk/images/125765-exam-standards-report-and-recommendations.pdf

Crisp, V., & Novaković, N. (2009a). Are all assessments equal? The comparability of demands of college based assessments in a vocationally related qualification. *Research in Post-Compulsory Education, 14*(1), 1-18. doi: 10.1080/13596740902717366

Crisp, V., & Novaković, N. (2009b). 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. doi: 10.1080/09500790902855776

Domyancich, J. M. (2014). The Development of Multiple-Choice Items Consistent with the AP Chemistry Curriculum Framework To More Accurately Assess Deeper Understanding. *Journal of Chemical Education*(91), 1347-1351.

Dunworth, K. (2008). A Task-Based Analysis of Undergraduate Assessment: A Tool for the EAP Practitioner. *Tesol Quarterly, 42*(2), 315-323. doi: 10.1002/j.1545-7249.2008.tb00126.x

Elwood, J., & Murphy, P. (2001). Tests, tiers and achievement: gender and performance at 16 and 14 in England. *European Journal of Education, 37*(4), 395-416.

Good, F. J., & Cresswell, M. J. (1988). Grade Awarding Judgements in Differentiated Examinations. *British Educational Research Journal, 14*(3), 263-281.

Greatorex, J., Shaw, S., Hodson, P., & Ireland, J. (2013). Using scales of cognitive demand in a validation study of Cambridge International A and AS Level Economics *Research Matters: A Cambridge Assessment Publication*(15), 29-37.

Hale, G., Taylor, C., Bridgemen, B., Carson, J., Kroll, B., & Kantor, R. (1995). A study of writing tasks assigned in academic degree programs. *ETS Research Report Series*(2), 1-61. doi: 10.1002/j.2333-8504.1995.tb01678.x

He, Q., Opposs, D., Glanville, M., & Lampreia-Carvalho, F. (2015). Assessing pupils at the age of 16 in England – approaches for effective examinations. *The Curriculum Journal, 26*(1), 70-90. doi: 10.1080/09585176.2014.944198

Hughes, S., Pollitt, A., & Ahmed, A. (1998). *The development of a tool for gauging the demands of GCSE and A Level exam questions*. Paper presented at the British Educational Research Association, The Queen's University, Belfast. http://www.cambridgeassessment.org.uk/images/109649-the-development-of-a-tool-for-gauging-the-demands-of-gcse-and-a-level-exam-questions.pdf

Johnson, M., & Mehta, S. (2011). Evaluating the CRAS Framework: Development and recommendations. *Research Matters: A Cambridge Assessment Publication*(12), 27-33.

Jones, D., & Tarr, J., E. (2007). An examination of the levels of cognitive demand required by probability tasks in

middle grades mathemetics books. *Statistics Education Research Journal*, *6*(2), 4-27.
Kegan, R. (1982). *The evolving self: Problem and process in human development*. Cambridge, MA: Harvard University Press.
Lave, J., & Wenger, E. (1991). *Situated Learning Legitimate Peripheral Participation*. Cambridge: CUP.
Long, R. (2016). GCSE, AS and A level reform (England) *Briefing Paper (House of Commons Library)*.
Maab, K. (2010). Classification Scheme for Modelling Tasks. *Journal für Mathematik-Didaktik*(31), 285–311. doi: DOI 10.1007/s13138-010-0010-2
Manwaring, M. (2006). The Cognitive Demands of a Negotiation Curriculum: What Does It Mean to “Get” Getting to YES? *Negotiation Journal*, *22*(1), 67-88. doi: 10.1111/j.0748-4526.2006.00086.x
Oates, T. (2013). Tiering in GCSE - which structure holds most promise? Cambridge: Cambridge Assessment.
Piaget, J. (1954). *The construction of reality in the child*. New York: Basic Books.
Pollitt, A., Ahmed, A., & Crisp, V. (2007). The demands of examination syllabuses and question papers. In P. Newton, J. Baird, H. Goldstein, & H. Patrick (Eds.), *Techniques for monitoring the comparaility of examination standards* (pp. 166-206). London: Qualifications and Curriculum Authority.
Shaw, S. D., & Crisp, V. (2012). An approach to validation: Developing and applying an approach for the validation of general qualifications. *Research Matters: A Cambridge Assessment Publication, Special Issue 3*.
Shute, G. E. (1979). Accounting Students and Abstract Reasonings: An Explanatory Study. Sarasota, FL: American Accounting Association
Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal, 33*(2), 455–488.
Tekkumru-Kisa, M., Stein, M. K., & Schunn, C. (2015). A Framework for Analyzing Cognitive Demand and Content-Practices Integration:Task Analysis Guide in Science. *Journal of Research in Science Teaching, 52*(5), 659–685. doi: 10.1002/tea.21208
Wenger, E. (Ed.). (1998). *Communities of Practice, Learning, Meaning and Identity*. Cambridge: CUP.
Wheadon, C., & Béguin, A. (2010). Fears for tiers: are candidates being appropriately rewarded for their performance in tiered examinations? *Assessment in Education: Principles, Policy & Practice, 17*(3), 287-300. doi: 10.1037/h0070288.
Yang, Y. (2009). How a Chinese teacher improved classroom teaching in Teaching Research Group: a case study on Pythagoras theorem teaching in Shanghai. *Mathematics Education: Exploring the Culture of Learning*(41), 279-296. doi: 10.1007/s11858-009-0171-y
Zohar, A., Schwartzer, N., & Tamir, P. (1998). Assessing the Cognitive Demands Required of Students in Class Discourse, Homework Assignments, and Tests. *International Journal of Science Education, 20*(7), 769-782. doi: 10.1080/0950069980200702
Zoller, U. L., A.; Nakhleh, M. B.; Tessier, B.; Dori, Y. J. (1995). Success on Algorithmic and LOCS vs. Conceptual Chemistry Exam Questions. *Journal of Chemistry Education*(72), 987−989.