Title: Measuring Critical and Creative Thinking Ability

Author: Ray Philpot, Australian Council *for* Educational Research

E-mail: Ray.Philpot@acer.edu.au

Abstract

Critical and Creative Thinking (CCT) is a core 21st Century competency. Is CCT ability a single, coherent construct that can be accurately measured, however? The Australian Curriculum, Assessment and Reporting Authority (ACARA) conceives CCT as a general capability that consists of four interrelated elements: Inquiring, Analysing, Generating and Reflecting. A study was carried out that tested whether CCT as defined by ACARA is a single construct that can be applied across school levels from Year 1 to Year 10. Trial data consisted of responses from 4,954 students in 48 Victorian Primary and Secondary schools to (various subsets of) 312 assessment items. Using Rasch measurement theory it was found that most items fit a single uni-dimensional model quite well. In this paper I outline the CCT construct, explain how the assessment instruments were developed and discuss characteristics of the final CCT scale.

Key words: Critical thinking, Creative thinking, Rasch measurement

1. INTRODUCTION

Definition of Critical and Creative Thinking

Critical and creative thinking has been defined in various ways.

- Critical thinking is reasonable reflective thinking directed at deciding what to believe or do. (Ennis, 1996)
- Critical thinking involves the ability to generate and evaluate knowledge, clarify concepts and ideas, seek possibilities, consider alternatives and solve problems. (ACARA, 2013)
- Critical thinking is a process which stresses an attitude of suspended judgment, incorporates logical inquiry and problem solving, and leads to an evaluative decision or action. (NCTE & IRA, 1996)
- Critical thinking is the ability to judge the plausibility of specific assertions, to weigh evidence, to assess the logical soundness of inferences, to construct counter-arguments and alternative hypotheses (Moore & Parker, 2012).
- Critical and creative thinking is skill at: generating ideas, clarifying ideas, and assessing the reasonableness of ideas (Swartz, 1998).
- Critical and creative thinking is thought of by some as synonymous with higher-order thinking. This is commonly typified as the three top levels of Bloom's Revised Taxonomy: Analysing, Evaluating, Creating (Anderson & Krathwohl *et al*, 2001).

The ACARA General Capabilities document (ACARA, 2013) states that "Critical thinking is at the core of most intellectual activity that involves students in learning to recognise or develop an argument, use evidence in support of that argument, draw reasoned conclusions, and use information to solve problems… Creative thinking involves students in learning to generate and apply new ideas in specific contexts, seeing existing situations in a new way, identifying alternative explanations, and seeing or making new links that generate a positive outcome."

Whilst it is recognised that *critical* thinking and *creative* thinking are not precisely the same thing, they are related and together are an integral part of thinking and learning. Furthermore, "critical and creative thinking can be encouraged simultaneously through activities that integrate reason, logic, imagination and innovation" (ACARA, 2013).

The study

A study was carried out by the author and a team at ACER to test whether critical and creative thinking as defined by ACARA is a single construct that can be applied across school levels from Year 1 to Year 10. The study was sponsored by the Victorian Department of Education and Early Childhood Development (DEECD) and the Victorian Curriculum and Assessment Authority (VCAA).

Assessment instruments were developed for the study with contexts for stimulus materials taken from a wide range of areas, including the sciences, humanities and arts: it is expected that skill in CCT is a part of successful thinking in any subject area. Detailed subject knowledge in any particular area was *not* assumed, as the aim was to test ability in CCT.

A few of the assessment tasks developed have been released for public viewing. These can be found on the new Insight Assessment Portal at http://www.insight.vic.edu.au/.

2. METHOD

The study was carried out in a series of six steps:

1. Development of an assessment framework based on ACARA's documentation for Critical and Creative Thinking (ACARA, 2013).
2. Development of the assessment tasks. Each task consisted of a series of short, independent items based around a theme, with a wide variety of contexts used across all tasks. The tasks covered a suitably wide range of ability of students in Years 2 through 10, with some tasks suitable for Year 1 with one-to-one administration. Detailed coding guides or scoring rubrics were written for each item.
3. Validation of the construct and assessment instruments by experts and teachers. This helped to ensure that CCT – and *only* CCT – was measured by the instruments.
4. Trialling of the assessment tasks in 48 schools with 4,954 students responding to the tasks.
5. Coding of the student responses by trained markers using the coding guides. This step included data entry and cleansing.
6. Performance of a series of analyses with the aim of constructing a single scale of critical and creative thinking ability, applicable from the end of Year 1 to Year 10, if possible.

The last step – item analysis – consisted of several processes, some performed iteratively: reviewing item (statistical/psychometric) characteristics; equating link items across year levels; recoding or deleting misfitting items; equating across test forms using link items; reviewing Differential Item Functioning; checking dimensionality, reliability and fit; and constructing the final scale.

The trial data consisted of responses to 312 items in 32 tasks, spread across 11 test forms, with vertical and horizontal links. A total of 48 schools participated with 4,954 students responding to (various subsets of) the items. The data set was analysed using a Rasch (one-parameter) partial credit model via the ACER *ConQuest* software (Wu *et al*, 2007).

In reviewing the items to determine whether they fit the model, the following characteristics were taken into account:

- Number of cases for this item
- Item-Rest and Item-Total Correlations
- Item fit – Weighted MNSQ
- Item Threshold(s) and Item Delta(s)
- Number of cases in each credit level, including missing data and % of total
- Point Biserial for each credit level
- Average Plausible Values for student ability at each credit level
- Standard Deviation for each Average Plausible Value

The Item Characteristic Curve (ICC) charts for each item in each test form at each year level were generated and inspected in turn and used in conjunction with the above item statistics in considering the psychometric behaviour of each item.

3. RESULTS

The main outcomes of the analysis are given in turn

Equating

An equating analysis across year levels for each test form was generated and inspected. A sample of the output produced is given in *Figure 1*. Items within the curved "railway tracks" have equated well.

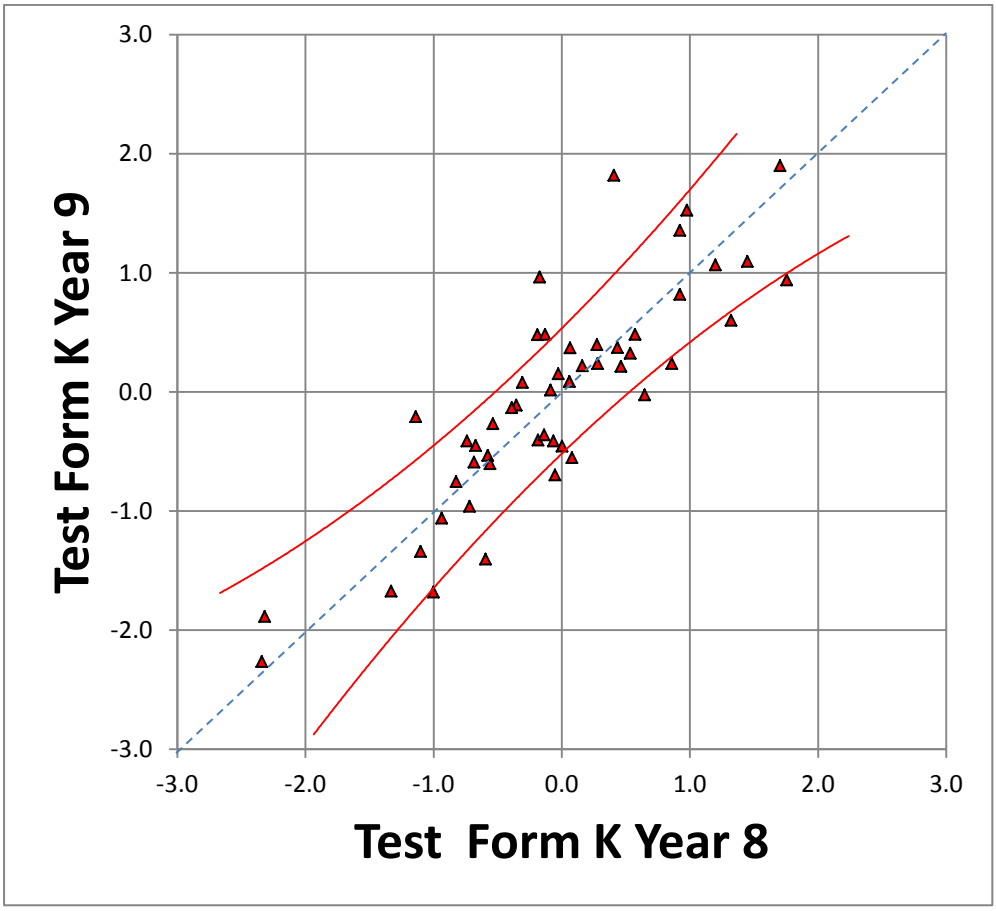

Figure 1 Example of a comparison of items in a test form across two year levels

Overall the equating worked reasonably well enabling difficulty values to be set for items in a test form independently of the age and ability of the student.

As a result of a review of the item and equating analyses, recodes were decided upon, possible deletions flagged and incorrect keys (in Multiple Choice items) were checked for.

Next, "between-test form" equating was performed. Equating the results from the different test forms was made possible by administering common items across two or more test forms and year levels. It was found that the relative locations (difficulties) of the common (or *link*) items were reasonably consistent across the assessment test forms and therefore that equating could be carried out. Where whole tasks appeared to be out of the expected average order of difficulty it was considered whether the target level of the task was appropriate and the level changed in some cases. Additionally, items that didn't fit well enough were re-coded or deleted. The remaining 285 items (containing 378 score points) had good psychometric properties and so could be considered for placement on a single (uni-dimensional) CCT scale.

Figure 2 gives an example of the variability in difficulty of link items. Differences can be ignored for the overall equating procedure, but can be a problem when it comes to attributing difficulty values to items *independently of which test form* they are in. This will be discussed further in a following section.

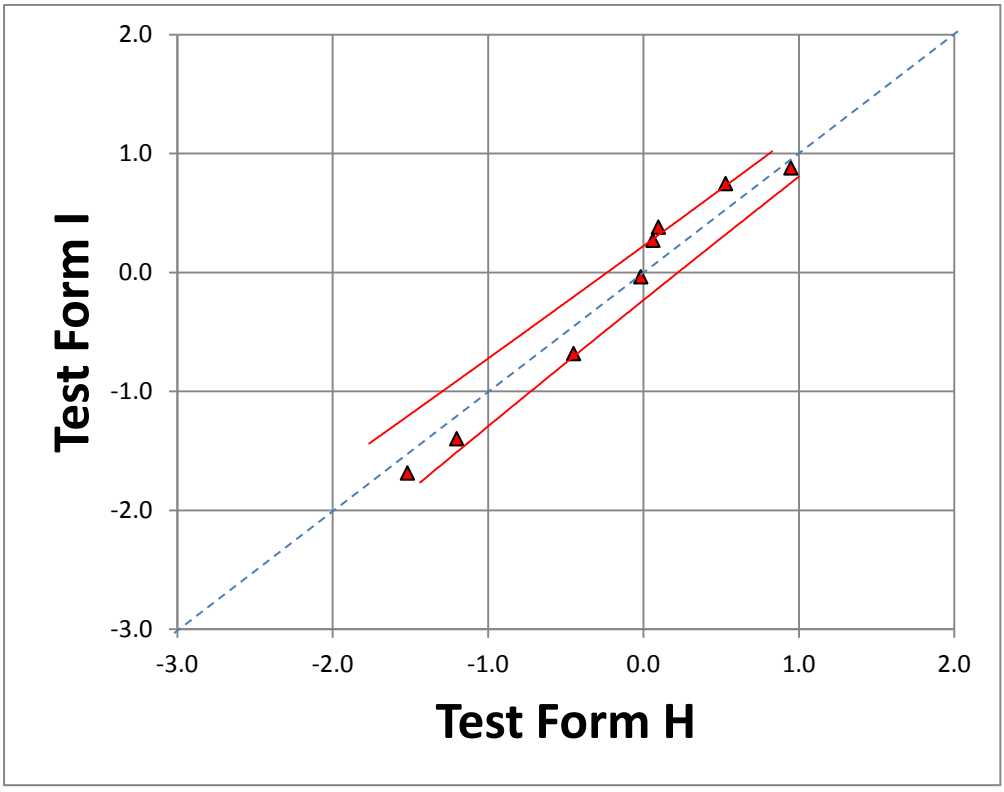

Figure 2 Between-test form equating – variations in link item difficulty

Differential Item Function (DIF)

Once the difficulty of an item is estimated independently of the ability of a student, the expectation is for that item's difficulty to remain the same (within error) no matter who attempts that item. If the difficulty of an item varies between different groups (after accounting for differences in ability) the item functions differently for different groups and hence displays Differential Item Functioning (DIF). DIF may be due to various factors. Year

level DIF, for example, could be due to the fact that students at a lower year level simply have not been taught the material required to answer an item.

DIF was found to be small overall with the impact of the DIF items on the measured ability of students being at most 0.06 logits in each test form. Similarly, using the equating results above, there were only small amounts of DIF across year levels.

Gender DIF

A Gender DIF analysis based on items in each test form was made. DIF was found to be small overall and the impact of the DIF items on the measured ability of students was at most 0.06 logits in each test form. Gender DIF for the entire item pool is illustrated in Figure 3, with items outside the error bands considered to have some DIF. There were 34 items favouring girls and 26 favouring boys.

It is interesting to note that with or without DIF items, girls tended to perform better in all test forms, in the range of 0.1 to 0.4 logits. It should be kept in mind, however, that since the sample was not random, one cannot generalise to the entire population.

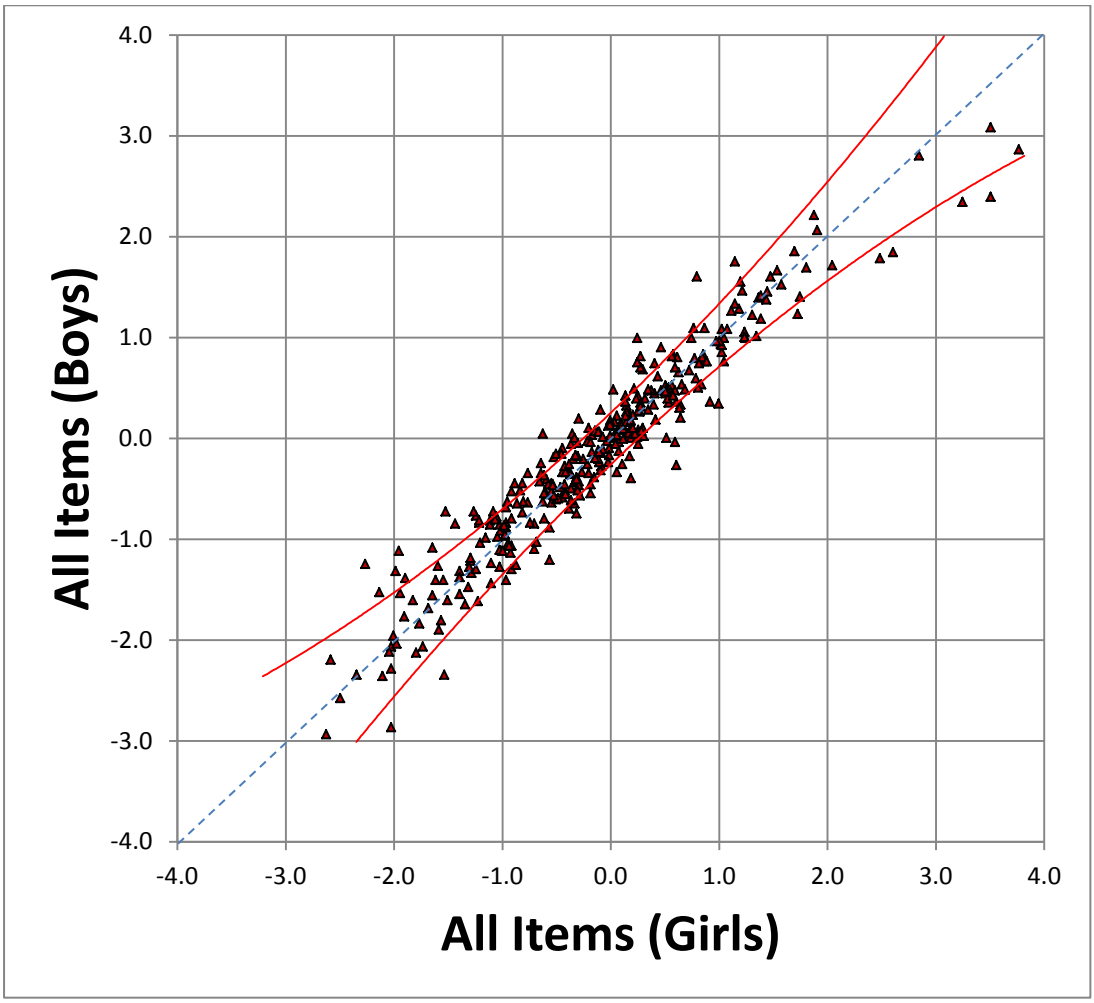

Figure 3 Gender DIF for all items

Region DIF

A School region DIF analysis based on all items was performed. Schools were categorised as inner suburban, outer suburban and rural. The results show that the impacts of DIF items on estimated abilities are very small (within 0.03 logits).

Joint item analysis

A joint item analysis was carried out using the remaining (285) items in an attempt to place them all on a single scale. As mentioned above in the equating analysis, some link items behaved significantly differently ($p<0.05$) in different test forms, so it is somewhat problematic to attribute a difficulty value for them. There were 49 such items. Among possible reasons for the difference in behaviour in these items are:

- the schools and classes were not randomly sampled and prior teaching in CCT was not controlled for;
- some items were very easy and others very hard;
- individual coder inconsistency;
- inter-coder differences; and
- measurement errors.

The difference in difficulties (across test forms) for these problematic items was mainly under 0.5 logits, and since measurement errors were as high as 0.37 logits per item, less than a dozen items are of any serious concern.

A way was sought to get a representative difficulty value for the inconsistent items, although one could always simply delete them from the pool. A two-step process was followed.

1. All common items in different test forms were treated the same, and a single item difficulty estimated for each item. [*A drawback of this step is that ignoring the equating results (particularly differences in year level of the students) and item differences across test forms (DIF) will result in inaccuracies.*]
2. Using the equating analysis results, for common items in different test forms with differing difficulties, the copy with difficulty closer to the difficulty found for the item in Step 1 is kept. [*This still does not fully account for DIF, however the comparison with the values from Method 1 will help to mitigate severe DIF effects.*]

This process resulted in a single CCT scale being constructed using a pool of 285 items.

Dimensionality

In the ACARA conception, CCT can be broken down into four elements: Analysing, Generating, Inquiring and Reflecting. Each item was placed into one of these four categories according to the main element drawn on by the item. A multidimensional IRT model was used to check whether a better fit could be obtained by treating the categories separately. The between-item analysis showed that the data had a better fit to the multidimensional model than the uni-dimensional model, as indicated by a statistically significant reduction in 'final deviance' in the model estimation. The correlations among the four item types are relatively strong, as seen in *Table 1*. Since the original uni-dimensional model fits quite well however, there is not enough reason to drop this in favour of a four-dimensional model.

	Analysing	Generating	Inquiring	Reflecting
Analysing				
Generating	0.79			
Inquiring	0.83	0.82		
Reflecting	0.74	0.77	0.79	

Table 1 Correlations between ACARA continuum elements

The item-type fit statistics are shown in *Table 2*. All weighted MNSQs are close to 1, a good result. The table shows that the Generating items had a weighted MSNQ slightly higher than the confidence interval of [0.95, 1.05], indicating Generating items had a slightly lower discrimination on average compared to the other three item types, although the magnitude of the misfit is not very large.

Item Type	UNWEIGHTED FIT				WEIGHTED FIT			
	MNSQ	CI_low	CI_high	T	MNSQ	CI_low	CI_high	T
Analysing	0.92	0.90	1.10	-1.64	0.94	0.94	1.06	-2.12
Generating	1.13	0.90	1.10	2.35	1.10	0.95	1.05	3.86
Inquiring	0.96	0.90	1.10	-0.82	0.98	0.90	1.10	-0.38
Reflecting	0.97	0.90	1.10	-0.67	0.97	0.93	1.07	-0.74

Table 2 Model fit statistics for four continuum elements

Reliability

Test reliability is generally defined as the proportion of the observed test score variance that is true variance. Values range from 0 to 1 with the higher the value the more reliable the instrument. In Item Response Theory, the reliability measure that best fits this description is separation reliability. For the joint analysis, before adjusting for DIF across test forms, the CCT separation reliability was 0.993. A Chi-square test of parameter equality gave a value of 45624.20 with 285 degrees of freedom and with a significance level of 0.000. In other words, the pool of items considered as a whole is highly reliable.

Expected A Posteriori / Plausible Value (EAP/PV) reliability describes biases in population parameter estimates: it measures how much variance in estimated ability data for a person is accounted for by the measurement model, averaged over all people tested. It is most valuable as an indicator of loss of precision due to the test design (Adams, 2005). For the joint analysis, before adjusting for DIF across test forms, the EAP/PV reliability was 0.902; for the final pool of 285 CCT items, the EAP/PV reliability was 0.890, both of which are good.

Typically the above reliability indices increase in size as the sample size increases. The number of items ranges from 29 to 51 for each trial test form and in all cases the values obtained were good.

Final scale

In short, the psychometric properties of the 285 items were good, the fit with a unidimensional model was acceptable, there was little DIF, the reliability of the test forms was good, and the items were believed by experts to measure CCT as defined by ACARA. Thus the final scale that was constructed (based on the joint analysis described above) is valid and applicable from the end of Year 1 through Year 10.

Figure 4 shows the difficulties of the items, ordered by *item score point difficulty* within each pair of year levels. That is, if an item has partial credit, the difficulty of the score point for partial credit is given separately from that for full credit. The scale is in logits: items located higher on the scale are more difficult than items located lower on the scale, and the scale is linear.

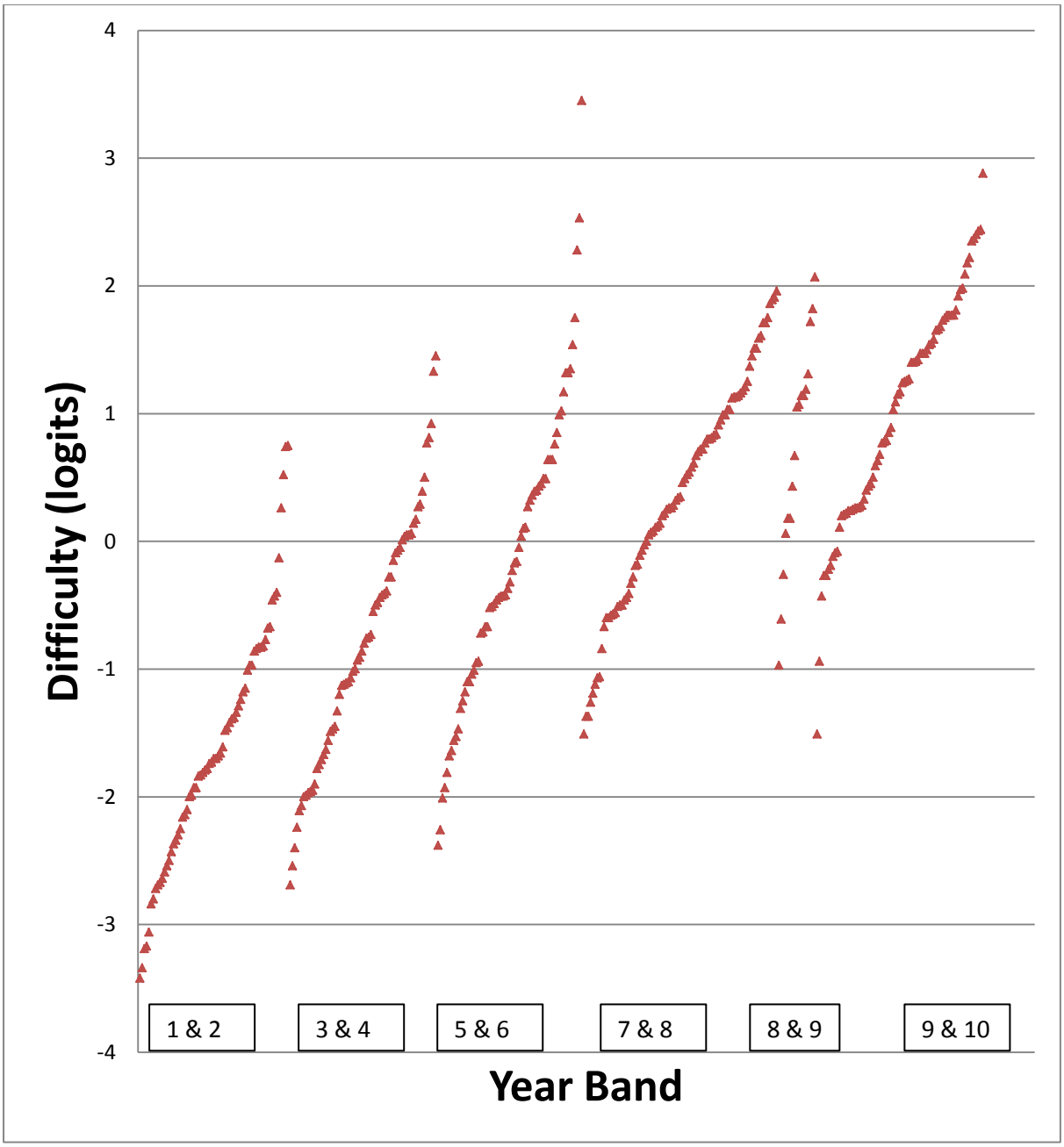

Figure 4 Item difficulty per score point by year band

It can be seen from the figure that there is substantial overlap in item difficulty between adjacent year-level pairs – this is by design, as the spread of CCT abilities of students in any year level is expected to vary greatly. In addition, it should be understood that a student with ability of say 1 logit will almost certainly not get all items with difficulty at or around 1 logit correct in a test. This is because the Rasch model is probabilistic, and a student whose ability is 1 has (by definition) a probability of 50% of getting an item of difficulty 1 correct.

4. CONCLUSION

The results obtained show that it is indeed possible to construct a single scale that measures critical and creative thinking ability in Primary and Secondary school students, and that the difficulty of test items can be calibrated on this scale.

The item statistics showed that on the whole there was good fit and discrimination. Using several different measures of test reliability the overall reliability ranged between 0.890 and 0.993, and for individual test forms between 0.713 and 0.988, indicating small amounts of measurement error.

Noting that "creative thinking" is not equivalent to "creativity", but is closer to the notion of generating possible solutions to a problem that is subject to constraints, there was no firm evidence in this study that "creative thinking" is a different skill from "critical thinking".

It is intended that the instruments developed in this study be used by teachers to measure CCT ability in students. The scale that has been constructed can be used to compare students and measure progress in individual students in the development of CCT ability. The scale does **not** say what constitutes an *acceptable* level of ability at any given year level, as the zero of the scale is essentially arbitrary. To achieve this would require an expert review of the scale in terms of the content of the items, their difficulties and their classifications in the assessment framework, with the aim of setting standards of achievement for each year level in terms of values on the scale. This is certainly a worthwhile exercise, but was beyond the scope of the current study.

5. REFERENCES

ACARA (2013). General capabilities in the Australian Curriculum. Retrieved from http://www.australiancurriculum.edu.au/generalcapabilities/pdf/overview

Adams, R.J. (2005) Reliability as a Measurement Design Effect. *Studies in Educational Evaluation, 31,* 162-172

Anderson, L., Krathwohl, D., et al. (eds) (2001). A Taxonomy for Learning, Teaching, and Assessing: a revision of Bloom's taxonomy of educational objectives. Allyn & Bacon, Boston, MA.

ATC21S (2013). What Are 21st-Century Skills? Retrieved from http://atc21s.org/index.php/about/what-are-21st-century-skills/

Butterworth, J & Thwaites, G. (2013). Thinking Skills – Critical Thinking and Problem Solving. Second Edition. Cambridge University Press, Cambridge, UK.

Ennis, R.H. (1996). Critical thinking. Englewood Cliffs, NJ: Prentice Hall

McCurry, D. (2013). Teaching critical thinking. *The Research Digest*, QCT, 2013, (9). Retrieved from http://www.qct.edu.au

Moore, B.N. & Parker, R. (2012). Critical thinking (10th ed.). New York: McGraw-Hill.

NCTE & IRA. (1996). Standards for the English language arts. Newark, NJ: IRA; Urbana, IL: NCTE
http://www.ncte.org/library/NCTEFiles/Resources/Books/Sample/StandardsDoc.pdf

Partnership for 21st century skills (2013). Retrieved from http://www.p21.org

Swartz, R. J. (1998). The design of infusion lessons. Retrieved from http://www.nctt.net/lessonsarticles.htm

Swartz, R, Costa, A, Beyer, B, Reagan, R, and Kallick, B. (2007). *Thinking-Based Learning*. Norwood, MA: Christopher Gordon Publishers

Wu, M. L., Adams, R. J., Wilson, M. R., Haldane, S.A. (2007). ACER ConQuest Version 2: Generalised item response modelling software Camberwell: Australian Council *for* Educational Research.