University of Wollongong

Research Online

Faculty of Social Sciences - Papers (Archive)

Faculty of Arts, Social Sciences & Humanities

1-1-2017

Evaluating the validity of the online multiliteracy assessment tool

Kellie Buckley-Walker

University of Wollongong, kelliebw@uow.edu.au

Jim S. Tognolini

University of Sydney, jimto@uow.edu.au

Lori Lockyer

University of Technology sydney, lori.lockyer@gmail.com

Ian M. Brown

University of Wollongong, ibrown@uow.edu.au

Peter Caputi

University of Wollongong, pcaputi@uow.edu.au

Follow this and additional works at: <https://ro.uow.edu.au/sspapers>

Part of the [Education Commons](#), and the [Social and Behavioral Sciences Commons](#)

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Evaluating the validity of the online multiliteracy assessment tool

Abstract

This study aims to assess the validity of the Online Multiliteracy Assessment for students in Years 5 and 6. The Online Multiliteracy Assessment measures students' abilities in making and creating meaning, using a variety of different modes of communication, such as text, audio and video. The study involved selecting two groups of students: the first group (n=19) was used in two pilot studies of the items and the second (n=299) was used in a field trial validating the functioning of the items and assessing the quality of the scale. The results indicated that the Online Multiliteracy Assessment has acceptable test-retest reliability; however, the fit to the Rasch model was less than ideal. Further investigation identified two important areas for improvement. First, the items assessing the higher order skills of synthesising, communicating and creating need to be more cognitively demanding. Second, some items need to be modified in order to improve their functionality.

Disciplines

Education | Social and Behavioral Sciences

Publication Details

Buckley-Walker, K., Tognolini, J., Lockyer, L., Brown, I. & Caputi, P. (2017). Evaluating the validity of the online multiliteracy assessment tool. *Australian Journal of Education*, 61 (3), 305-327.

Introduction

The pervasiveness of technology requires people to be able to make and create meaning from a variety of modes of communication. This is essential for social interactions, working, leisure activities and learning. Students today, from pre-school to university, interact with information presented in a range of modes in which they can make meaning of their world and their specific learning tasks. This technology-integrated way of making meaning has led to calls from researchers and educators to broaden the definition of literacy (Baguley, Pullen, & Short, 2010).

The term ‘Multiliteracies’ has been used to represent a societal context that comprises multiple forms of media and communication and emphasises “cultural and linguistic diversity” (New London Group, 1996, p. 63). Other researchers (e.g., Baguley et al., 2010; Burke, Butland, Roberts, & Snow, 2013; Unsworth, 2002) support these concepts by arguing that today’s learners will not only require traditional forms of literacy, but also, a broader skill set. This is required if learners are to make and create meaning from the various forms of communication available to them in the multimodal environment in which they live and learn. However, due to continuing research into literacy, an agreed definition of multiliteracies is difficult to locate. For the purposes of this article, multiliteracies has been defined using two concepts: *that language is used to make meaning in different ways in different contexts and meaning is made through the use of different modes and combinations of modes* (Dobson, S, 2003; Kress, 2003).

To further develop the idea of multiliteracies, we propose a conceptual model which outlines and describes the skills associated with multiliteracy. The conceptual model (see Figure 1) is supported by the technology-facilitated environment in which students live and learn, and requires them to make meaning from a variety of modes of communication. The eight skills,

as detailed in Figure 1, which comprise the conceptual model, are based on Bloom's original taxonomy plus other more recent revisions (Anderson & Krathwohl, 2001; Bloom, 1956; Churches, 2009; Turcsányi-Szabó, 2012). Each skill comprises of a continuum ranging from lower level to higher level, with the higher level demonstrating a more critical approach.

Insert Figure 1 about here

From theory and research (e.g., Bloom, 1956; Starkey, 2011), the skills that are presented in the model will vary in terms of their difficulty. Difficulty relates to how cognitively demanding the students find each of these skills. For example, those skills that are less demanding, such as locating, the students will find easier and obtain higher scores in assessments. However, those skills which are more demanding, such as communication, the students will find more difficult and obtain lower scores in assessments. These skills could be placed on a developmental continuum, with less cognitively demanding items placed further to the left and more cognitively demanding items placed further to the right.

In the Australian education context, the national curriculum places an emphasis on multimodal learning and the new conceptions of literacy in all of its curricula. This focus encourages teachers to devise lessons using a variety of modes of communication and for students to make meaning from these modes. However, what needs to be answered is to what extent the students can make meaning from a variety of different modes of communication.

Addressing this question requires teachers to assess students' development of the skills associated with multiliteracies. Current assessment practices that use paper-based formats are too narrow and limited to cater for the technologically rich environments in which students learn today (e.g., Brown, Lockyer, Caputi, & Tognolini, 2010; Dawson & Siemens, 2014; Jewitt, 2003; Johnson & Kress, 2003; Kalantzis, Cope, & Harvey, 2003). However, there is

no comprehensive assessment tool that caters for the complexity of learning designs presented to contemporary students.

The current study

The study presented in this article was part of a research project that developed and validated an instrument which used a variety of modes to assess multiliteracy skills. The online multiliteracy (o-Mlit) assessment was developed to measure students' abilities to make meaning from text, sound, image and video (i.e., multiliteracy skills) in an online environment. As the focus of the o-Mlit was to assess students' multiliteracy skills, the topics, or content, on which the skills were assessed, could be variable. The three topics chosen were identified as being of interest to students through focus groups with Year 5 and 6 school teachers. The chosen topics included - 'Jump from Space' (JS), 'Social Media' (SM) and 'Health and Nutrition' (HN).

The items for the o-Mlit were developed through an iterative process involving feedback from expert -reviewers and teacher groups. The items were based on multiliteracy research and the conceptual model of multiliteracy (see Appendix A for example items). The structure of the o-Mlit is outlined in Table 1. Content validity was informed by what the item writers and expert reviewers thought about the items in relation to how well the items were assessing the construct of multiliteracy.

Insert Table 1 about here

The o-Mlit assessment is structured in three sections with a total of 42 items. All topics (JS, SM and HN) were assessed in section one. In section 2 students selected one of the three topics to be the focus. Section 3 asked students to plan and develop a presentation based on the topic selected in section 2. The presentation (e.g., powerpoint with embedded videos, word document with inserted images) that the students create is uploaded to the o-Mlit site.

This article reports on a study undertaken to examine the construct validity of the o-Mlit instrument, to identify whether it is measuring what it is purporting to measure. The validity was assessed by focussing on two types of evidence – response process and internal structure (Messick, 1989). The response process focussed on the actions and thought processes of those students undertaking the assessment. The internal structure focused on the reliability and the factor structure of the assessment. The response process and internal structure were investigated in two steps; small pilot studies and then a larger field trial.

Pilot studies

Participants

Two pilot studies were conducted. Pilot study 1 was based on the initial version of the o-Mlit while pilot study 2 was based on the revised version of the o-Mlit. For pilot study 1, nine Year 5 and 6 students from three public schools in New South Wales (NSW) were selected by their classroom teachers. The selection of students was based on the teachers' knowledge and judgements about the students' ability to represent a range of literacy levels and be confident enough to give their opinions about the o-Mlit to the researchers.

The average age of the nine students was 11.73 years (range = 11.08 yrs a – 12.57) with the majority being female (55.56%) and in Year 6 (55.56%). These students were provided with information sheets and consent forms for themselves and their parents. On the first day of testing, consent forms were collected and students' understanding of the research process was checked and any misconceptions clarified.

For pilot study 2, ten students were selected through a snowballing technique. The snowball sampling was used as the study was conducted during school holidays when access to students through schools was not available. The snowball sampling began with an email to employees within the local university. Information sheets and consent forms for the students and their parents were sent via email. The average age of the ten students at the time of the

study was 11.08 years (range = 9.68 yrs – 12.55 yrs a) with the majority being female (70.00%) and in Year 6 (60.00%). Again on the first day of testing, consent forms were collected and students' understanding of the research process was checked and any misconceptions clarified.

Method

The o-Mlit was administered to of the 19 students in order to ascertain their reactions to the items (e.g., requiring clarification, reason for selecting a distractor in a multiple choice item). Assessing the response process, that is the actions and thought processes of those students undertaking the test, can highlight what the students actually do during the assessment and how that matches with the constructs being assessed (Cook, Cook, & Beckman, 2006).

For each pilot study, students' reactions to the items in the o-Mlit were captured using audio-recorded 'Think Aloud Protocol' (TAP) (Ercikan et al., 2010) and screen-capture technology (e.g., Barmby, Harries, Higgins, & Suggate, 2009). Students were asked to speak about everything they were thinking while completing the task. It was suggested that they acted as if they were alone in a room speaking to themselves (e.g., Camps, 2003).

Students were also provided with a description of how their mouse movements and clicks on the screen as well as their voice would be recorded using 'Camtasia Studio 8' software.

Students were provided with a demonstration of the TAP and Camtasia, and were then able to practice using the protocols and technology.

For pilot study 1, students completed the study one day a week over four weeks, for a maximum of 1 hour. Students completed section 1, section 2, section 3 – planning and section 3 – creating. This process ensured that the impact on the student was reduced as well as increasing the flexibility for the classroom teacher, in regards to the timing of the study. In pilot study 2, students were only asked to work through section 1 and 2 due to the majority of

amendments occurring in these sections. To reduce the impact on the family and students, testing was completed in one session in the second pilot study, lasting a maximum of 1.5 hours.

Analysis

When examining item validity in terms of response processes, identifying the link between the item and how the students respond to that item was important. Therefore, the thoughts and questions of each student about each item, as well as his/her movements on screen, were transcribed and analysed by the researchers.

A two-stage process based on grounded theory was used to identify codes to analyse the students' responses to the items (Charmaz, 2006). Initial coding was the first step in the process and focussed on a few students' responses to each item of the o-Mlit. This highlighted a number of issues students encountered when using o-Mlit. Focussed coding was then completed which applied the initial codes to all students' responses. The codes developed from the focussed coding were subsequently grouped into two themes, namely 'test items' and 'technical issues'. Finally, the researchers discussed the two themes and identified ways to improve the items, source material and/or layout of the o-Mlit assessment. This process occurred for both pilot studies.

Results

For pilot study 1, students' responses to items were coded as shown in Table 2. As a consequence of the results, the researchers changed the wording of some of the questions (e.g., changed 'adolescent' to 'teenager'); changed some of the source material (e.g., made the access to websites clearer; changed from text-link to thumbnail-link); changed some of the response formats (e.g., introduced a different text box for each of the required responses asked within a question) and reformatted some of the items to make them more similar. It became clear that some of the source material was quite complex for the less proficient

students. However, the researchers decided not to change the more complex source material because the material was intended to be as authentic as possible.

Insert Table 2 about here.

One major change that did occur as a consequence of the results of the first pilot study was to include an introduction page. The introduction page was designed to enable students to practice various item formats that would be used in the o-Mlit. A second major change was the inclusion of audio for each question. This gave the students an option of listening to the question rather than just reading it. The ability to listen to the question was of particular importance for those students with lower literacy attainment, as it enabled them to have an understanding of what the question was asking. Once these changes were made to the o-Mlit, further testing was carried out on the revised o-Mlit in pilot study 2.

Data from pilot study 2 were analysed as shown in Table 2. As a result of this analysis some minor modifications were made to the revised oMlit and involved moving the question audio button to the right of the question, adding an extra instructional phrase (e.g. 'Using the information below'), highlighting (e.g., job) or removing a word (e.g., helmet) in a question, and decreasing the time to complete Section 1 from 45 minutes to 30 minutes. These modifications led to the development of the final version of the o-Mlit which was used in the next step, the field trial.

Discussion

The purpose of the pilot studies was to investigate the response process of the o-Mlit.

Evidence for the response process focused on what the students did and thought during an assessment and the extent to which this was in accord with the construct being assessed. This evidence was used to amend those items where the students were focussing on irrelevant

material which distracted from the construct being assessed (Miller, Linn, & Gronlund, 2013) thus lowering the validity of the item.

In both pilot studies the theme of 'test items' identified the extent to which the item was assessing the targeted skill. For example, by the highlighting or removal of a word in either the question or instructional information made the item clearer to students and hence focussing on the skill being assessed which also added to the evidence for the content validity as mentioned previously (Miller et al., 2013).

The 'technical issues' identified the ability of the students to navigate through the o-Mlit as well as the consistency of the layout between items and sections. For example, during pilot study 1 some students did not know how to select an answer to the item (e.g., drop down menus), move to the next item or play the audio. This was not surprising as many Year 5 and 6 students had not had previous experience with online assessment (ThinkLink Consultancy, 2014). Still, it detracted from the validity of the test because it was not related to the construct of multiliteracy. Therefore, these technical issues had to be addressed in order to improve the validity of the o-Mlit. This was done by adding an 'Introduction' page where student could practice interacting with the items. The introduction page also contained videos explaining the types of responses (e.g., multiple choice), the types of source material, the layout of the items. The inclusion of the introduction page enabled the students to focus on the items rather than what button to press or how to get the audio to start (Doorey, 2014).

The data gathered from pilot study 2 indicated that many of the technical issues identified in the pilot study 1 had been resolved. The evidence gathered from the pilot studies also suggests that the relationship between the construct of multiliteracy and the thought processes and actions carried out by the students were highly related. This relationship was demonstrated by students' expressing greater clarity in responding to the items. In other

words, students had a better understanding of what the items required of them, did not focus on distracting information and clearly identified the source material. This increased alignment between the construct being assessed and students' responses indicated readiness for the next phase of testing.

Field trial

Participants

Two hundred-and-ninety-nine students from 12 schools from rural and metropolitan NSW participated in the field trial. Each student received information sheets and consent forms for themselves and their parents which were returned to the school and collected and checked by the researcher for completeness. The students had an average age of 10.97 years (range = 9.58 years – 12.50 years) with the majority being male (51.66%) and in Year 6 (53.73%).

The o-Mlit was administered on three occasions over 6 months, 2014. The average time between occasions was 2.6 months (range of 2.1 to 3.2 months). The number of participants decreased over each testing occasion (occasion 1 – 291; occasion 2 – 285; occasion 3 – 275). The reasons for this decrease in participant numbers included student absences (7.69%), leaving the school (2.68%) and withdrawal from the study (1.34%). Seven students joined the study at occasion 2 which meant that the total decrease in participation from occasion 1 to occasion 3 was 6.7 per cent. However, the number of complete data sets increased over the three testing occasions (occasion 1 – 184; occasion 2 – 224; occasion 3 – 271). The low response rate of 63.23 per cent of the total sample at occasion 1 was due to a technical issue that occurred when the participants uploaded their presentations, however, this was rectified for occasion 2. The ratio of participants to items was 7:1.

Method

The purpose of the field trial was to investigate the internal structure of the o-Mlit. Students were tested with the o-Mlit on three separate occasions, in order to determine the reliability of

the o-Mlit. Testing was conducted over four days to reduce the amount of fatigue experienced by students which may have impacted the results and to provide more flexibility to participating schools in regards to the timing of the testing.

On the first day of the field trial, the researcher explained the purpose of the study, answered questions, helped the students log onto the o-Mlit site and assisted them to complete the demographic questionnaire. This process allowed any technical difficulties to be identified and addressed prior to the start of the test.

The testing procedures for the next four days are shown in Figure 2. The videos played on days one, three and four explained to students how they were to answer the items (e.g., multiple choice, dropdown boxes, and moving images), play audio and video, listen to the questions and move between items. Each day of testing, students logged on to the o-Mlit site and click on the section to be completed on that day. For example, Day 1 students clicked on 'section 1' on Day 3 students clicked on 'section 3 plan'. The research assistant (RA) timed how long the students had to complete the section. Once time was up the students logged off the o-Mlit site. These procedures were followed on the three testing occasions.

Insert Figure 2 about here.

Analysis

Rasch Theory (Rasch, 1960/1980) was used to determine the extent to which the items of the o-Mlit comprised a single construct (unidimensionality); whether the items functioned in a manner that was consistent with the expectations of the Rasch Model; whether the items showed any form of differential item functioning (DIF) for gender; and whether the marking rubrics had functioned appropriately. The Rasch Unidimensional Models for Measurement (RUMM2030) program was used to analyse the data.

An analysis of variance (ANOVA) was conducted to identify any differences between the overall mean scores that students obtained on each of the three testing occasions (Field, 2009). To identify if the mean scores in each skill assessed and the overall scores were consistent over the three testing occasions Pearson correlation coefficient (test-retest correlations) were conducted (Field, 2009). Both the ANOVA and Pearson correlation coefficients were conducted using IBM SPSS Statistics 21 (IBM Corp, 2012), at an alpha level of .05.

On the basis of the results obtained from the field trial, a focus group comprised of classroom teachers was conducted. The purpose of this focus group was to identify possible reasons for the items not functioning as expected. The focus group participants were given information (e.g., graphs identifying bias) about the items and asked to suggest reasons as to why the nominated items may not have performed as expected. The field notes were analysed using grounded theory (Charmaz, 2006) to identify possible reasons for the items' non-performance.

Results

The students could choose the topic on which they would focus at each testing occasion for sections 2 and 3. The most popular topic for occasion 1 was Social Media (SM: $n=125$), then Health and Nutrition (HN: $n=85$) and then Jump from Space (JS: $n=80$). On occasions 2 and 3, HN was the most popular topic ($n=125$ and 112 , respectively), followed by SM ($n=93$ and 92 , respectively) and finally JS ($n=63$ and 67 , respectively). The process for analysing the internal structure of the o-Mlit was a top-down approach by first investigating the o-Mlit as a task, then the skills and finally the individual items.

Overall Task. The test-retest reliability of the o-Mlit over time was - occasion 1 to occasion 2 - $r(143) = .78, p < .00$; occasion 2 to occasion 3 - $r(192) = .77, p < .00$. These results indicated that the scores that the students achieved on the o-Mlit are relatively consistent for

each testing occasion. The mean scores that the students obtained on each of the three testing occasions (occasion 1 – 51.7%, SD - 11.6; occasion 2 – 52.3%, SD – 14.0; occasion 3 – 57.1%, SD – 14.3) did not vary significantly, $F(2, 524) = 0.11$, $p = .99$. Therefore, there was no significant difference between the student ability estimates across the three testing occasions. The chi-square goodness-of-fit index was statistically significant, $\chi^2(509) = 1064$, $p < .000$, indicating a poor fit to the Rasch model, suggesting that the o-Mlit was assessing more than a single factor. However, it may also have indicated that a high residual covariance among the items and skills which increased the likelihood of a significant result as well as the probability of a poor fit to the Rasch model.

Skills. The students varied in their performance on the skills assessed within the o-Mlit across the three testing occasions. The test-retest correlations ranged from applying $r(271) = .08$, occasion 1 to occasion 2; to locating $r(269) = .65$ from, occasion 2 to occasion 3 (see Table 3). The internal consistency of each of the skills ranged from relatively poor to relatively good over the three testing occasions, refer to Table 4. For example, applying skill at occasion 1 had a Cronbach's $\alpha = .02$ (poor) where communicating skill at occasion 2 had a Cronbach's $\alpha = .78$ (good). For all skills comprising three or more items, the reliability would increase if at least one item was deleted. These results may suggest that some of the items in each scale may be irrelevant, ambiguous or heterogeneous; this will be investigated in the item analysis section.

Insert Tables 3 and 4 about here.

It was expected that the items that were more cognitively demanding would be further to the right on the developmental continuum (relatively more difficult). Items which assessed the skill of locating would be the least cognitively demanding and, as such, would be to the left hand end of the developmental continuum. The next most demanding items would be those

which assessed understanding, applying, analysing, evaluating, synthesising, creating and communicating being the most difficult (Anderson & Krathwohl, 2001; Bloom, 1956; Churches, 2009; Turcsányi-Szabó, 2012). However, the data did not support this expected ordering of the skills.

Figure 3 shows the difficulty of each skill based on the mean of the item difficulties which assessed each of the skills over the three testing occasion. This figure shows that the order based on cognitive demand (difficulty) followed the expected order up until evaluating. The skills of synthesising, creating and communicating were not in the predicted order, and are located between understanding and applying in terms of their relative difficulty. In other words the students did not find these skills more cognitively demanding than those assessing the less cognitively demanding items. This finding suggests that the skills may need to be clarified further in terms of their definition; or the items are not assessing these higher order skills; and/or, the marking rubrics may not be functioning in the way that they were intended to do.

Insert Figure 3 about here.

Items. Table 5 shows the items of the o-Mlit arranged in difficulty order based on their average difficulty across the three occasions. It can be seen that item '1a' is the easiest and item '28' is the most demanding (or hardest) item. Table 5 also shows a number of fit statistics for each of the items on each of the occasions. The Residual-fit Statistic and the Chi-square Statistic give an indication as to how well each item fits the Rasch theory which governs the construction of the scale underpinning the test.

Items with residual-fit statistics greater than +2.5 have discriminated less than what the model would have predicted; those items with fit-residuals less than -2.5 have discriminated better than would have been predicted by the model (RUMM Laboratory Pty Ltd, 2009).

From a validity point of view, the question is why are the items under or over discriminating? Chi-square statistics coupled with the probability of fit to the Rasch model gives an indication of how the item fits the Rasch Model. If the probability of fit to the model is less than 0.05, then the question is why is this item not functioning in accord with the model?

Insert Table 5 about here

Table 5 shows that 14 items (in bold) have violated one or more of the fit criteria. Reference to Table 5 shows that item 19 on occasion 1 had the largest positive fit-residual (+4.3). The Item Characteristic Curve (ICC) for this item is shown in Figure 4 (a). The ICC gives an indication of what a positive fit-residual looks like in terms of the theory, the solid line on the graph. What is actually occurring is indicated by the dots, it shows that the less proficient students have done better on this item than would have been expected and the more proficient students have done a bit worse (the item under-discriminates). Similarly, the ICC for item 10 on occasion 3 (see Figure 4 (b)) shows an item with a relatively large negative fit-residual (-5.2). The ICC shows that the more proficient students have done better than predicted and the less proficient students have done worse (the item over-discriminates). The ICCs for item 14 (see Figure 5) on each of the three occasions shows how this item has varied in regards to the fit to the model over the 3 occasions.

Insert Figure 4 and 5 about here

There are six items (24, 26, 27, 28, 32 and 36; refer to Table 5) that are not functioning correctly due to issues related to the category probability and threshold disorder. The category probability indicates that scoring categories are not behaving as intended, as one or more scores (e.g., a score of 2 out of a possible 4) are not likely to occur. Threshold disorder again indicates that scoring categories are not behaving as intended, in that the probability of

scoring a higher mark is easier than scoring a lower mark; for example, scoring 3 out of 4 is easier than scoring 2 out of 4.

Three items (23, 28 and 31; refer to Table 5) are displaying DIF which occurs when the items are not functioning the same way for different groups, for example, gender, socioeconomic status or ethnicity. In this case DIF has occurred between students doing the different topics – JS, HN and SM. This suggests that these items are not functioning in the same way, which contradicts the intentions of the items writers.

The focus group were asked to consider the items that did not perform as expected (those items in bold in Table 5). The focus group participants identified possible reasons for non-performance: source material (e.g., text in the source material too difficult), screen layout (e.g., scrolling through the item), differences between topics (e.g., image in one topic easier to interpret than the other two topics), marking guidelines (too many options in the marking criteria), and questions (e.g., not explicit enough to direct the students). Although, there were 14 items that did not perform as expected, some of these items may have had more than one reason for the non-performance.

Discussion

Overall, the o-Mlit had good test-retest reliability. Further evidence of the consistency of performance on the o-Mlit was also provided by the non-significant results of the ANOVA for the mean student scores. The o-Mlit had a relatively poor fit to the Rasch model. This finding implies that the items of the o-Mlit are not functioning in the way that was expected when the test was being constructed. In order to investigate potential reasons for the misfit, the skills and items of the o-Mlit were investigated further using a focus group comprised of classroom teachers who interrogated the functioning of the items using the item analysis provided by the RUMM program.

When constructing the test, the item writers predicted that the skills represented various levels of cognitive depth which were largely in accord with Bloom's Taxonomy (e.g., Anderson & Krathwohl, 2001; refer to Figure 3). When the items were constructed the expectation was that the mean difficulty of the items assessing locating would be lower than the mean difficulty of the items assessing applying. The result supported the original expectations, except for the higher order skills of synthesising, creating and communicating. This finding was unexpected because it has been suggested (Anderson & Krathwohl, 2001; Ash, Clayton, & Atkinson, 2005; Kimber & Wyatt-Smith, 2010) that the cognitive demand of the higher order skills (e.g., creating) is greater than that of the lower order skills (e.g., understanding).

One reason for the unexpected result could be that the items comprising these skills are not assessing the skills adequately and/or it could be that the marking rubrics may not be functioning in the way that they were intended to do, as developing items which assess higher order skills is difficult (Lemons & Lemons, 2013). Another reason could be that the students have internalised these skills to a level where they are no longer assessing higher order thinking. If this is true then it would suggest that the initial construct does not reflect the current reality.

The focus group reviewed the higher order items using the analysis to determine whether it was the construct itself that needed to change as a result of the data or whether the items and marking rubrics were not adequately assessing the higher order skills that they had been written to assess. A number of reasons were identified by the focus group for the differential performance of the non-fitting items, such as DIF between the topics. For example, students who completed the synthesising item using one topic stimulus (e.g., SM) did significantly better than students completing the same item in another topic (e.g., JS). This contextual effect influences the validity of the test and as such needs to be addressed, particularly if the

original item writers were of the opinion that the items difficulties should not vary with context. Therefore, the homogeneity between topics needs to be improved (Griffin & Care, 2014). Another synthesising item has issues related to the marking rubric. In this case there were too many options within one grade for the markers to be aware of when marking. This increase in complexity for the marker, leads to a decrease in accuracy; therefore, influencing the validity of the decision made from the item (Griffin & Robertson, 2014).

The overall view of the focus group was that the test has assessed the lower order skills of multiliteracy quite well; however, the challenge is to construct items that are better assessing the higher order skills under test conditions in a multimodal environment. This challenge begins with how the higher order skills are defined, as a broad definition may increase the heterogeneity of the item.

There are a variety of terms used that are described as higher order skills such as: transfer, critical thinking and problem solving (Brookhart, 2010); inventive thinking skills, which includes 'higher order thinking and sound reasoning' (Abdullah & Osman, 2010); higher order reading skills (Afflerbach, Cho, & Kim, 2015); articulated learnings (Ash et al., 2005); problem-solving taxonomy (Barak, 2013). However, these terms all include skills associated with each other and multiliteracy for example - critical thinking skills include applying, analysing, evaluating and communicating; higher order thinking skills involves locating or recalling, applying, analysing, evaluating and synthesising. What is required is a clear definition of the skills underlying these various higher order skills in order to develop sharper and more cognitively demanding items which tap into these constructs (Fitzpatrick & Schulz, 2015).

One of the limitations of this study is the desirability to assess students' computer skills to investigate how that impacts their ability to navigate the online assessment. Having poorer

computer skills will negatively impact the students' ability to navigate and complete the online assessment, as observed in the pilot study; therefore, future studies will take students computer skills into account. Future studies will also increase the numbers of participating students to increase the generalisability of the findings. Through improvement of poor performing items, increasing the cognitive demand of the items assessing higher order skills (synthesising, communicating and creating) and further testing this will improve the overall functioning of the o-Mlit. Future studies will also focus on the relations to other variables (e.g., literacy skills) as well as the information in terms of diagnostics that the o-Mlit can provide.

Conclusion

The o-Mlit is based on a conceptual model of multiliteracy skills with the purpose of assessing each of these skills to provide teachers with knowledge about each student's performance. As outlined in the introduction, where the purpose of assessment is to make inferences about student learning it is important that the assessment data on which these inferences are made are valid. The objective of this study was to examine the construct validity of the o-Mlit by obtaining evidence regarding response processed and the internal structure of the instrument.

In general, the evidence presented in this article indicates that the o-Mlit is assessing the lower order skills effectively. However, some evidence suggests that the items designed to assess the higher order skill, namely, synthesising, communicating and creating have not functioned as expected. This may be due to either the fact that items are not assessing the higher order skills or that a clearer definition of what these higher-order skills mean in the context of multiliteracies in order to write items which better tap into these skills.

In addition, the analyses showed that some of the items did not function in accordance with the measurement model that has governed the construction of the assessment and would need to be rewritten. However, the analyses presented in this article indicate that a big step forward has been made with the development of the o-Mlit towards assessing multiliteracies which are an important element in mastering the multiple forms of media and communication in the 21st century.

References

- Abdullah, M., & Osman, K. (2010). *21st century inventive thinking skills among primary students in Malaysia and Brunei*. World Conference on Learning, Teaching and Administration Papers, 9, 1646–1651. Retrieved from <https://doi.org/10.1016/j.sbspro.2010.12.380>
- Afflerbach, P., Cho, B.-Y., & Kim, J.-Y. (2015). Conceptualizing and Assessing Higher-Order Thinking in Reading. *Theory Into Practice*, 54(3), 203–212. Retrieved from <https://doi.org/10.1080/00405841.2015.1044367>
- Anderson, L. W., & Krathwohl, R. (2001). *A taxonomy for learning, teaching and assessing: A revision Bloom's taxonomy of educational objectives*. New York: Longman. Retrieved from <http://ereadings.uow.edu.au.ezproxy.uow.edu.au/andersonl1.pdf>
- Ash, S. L., Clayton, P. H., & Atkinson, M. P. (2005). Integrating reflection and assessment to capture and improve student learning. *Michigan Journal of Community Service Learning*, 11(2), 49–60.
- Baguley, M., Pullen, D. L., & Short, M. (2010). Multiliteracies and the new world order. In D. R. Cole & D. L. Pullen (Eds.), *Multiliteracies in motion: Current theory and practice* (pp. 1–17). New York: Routledge.

- Barak, M. (2013). Teaching engineering and technology: cognitive, knowledge and problem-solving taxonomies. *Journal of Engineering, Design and Technology*, 11(3), 316–333. Retrieved from <https://doi.org/10.1108/JEDT-04-2012-0020>
- Barmby, P., Harries, T., Higgins, S., & Suggate, J. (2009). The array representation and primary children's understanding and reasoning in multiplication. *Educational Studies in Mathematics*, 70(3), 217–241. Retrieved from <https://doi.org/10.1007/s10649-008-9145-1>
- Bloom, B. S. (1956). *Taxonomy of educational objectives : The classification of educational goals..* New York : D. McKay Co., 1956-1964. Retrieved from <http://ezproxy.uow.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cat03332a&AN=uow.b1184472&site=eds-live>
- Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom.* Alexandria, Va. : ASCD, c2010. Retrieved from <http://ezproxy.uow.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cat03332a&AN=uow.b2160773&site=eds-live>
- Brown, I. M., Lockyer, L., Caputi, P., & Tognolini, J. S. (2010). *Multimodality, multiliteracy and visual literacy: Where does assessment fit?* Research Online. Retrieved from <http://ezproxy.uow.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ir00653a&AN=wuro.hbspapers.3245&site=eds-live>
- Burke, A., Butland, L., Roberts, K., & Snow, S. (2013). Using multiliteracies to 'rethink' literacy pedagogy in elementary classrooms. *Journal of Technology Integration in the Classroom*, 5(1), 41–53.
- Camps, J. (2003). Concurrent and retrospective verbal reports as tools to better understand the role of attention in second language tasks. *International Journal of Applied Linguistics*, 13(2), 201–221. Retrieved from <https://doi.org/10.1111/1473-4192.00044>

- Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. London: SAGE Publications Ltd.
- Churches, A. (2009). Bloom's digital taxonomy v3.01. Retrieved from <https://edorigami.wikispaces.com/file/view/bloom%27s+Digital+taxonomy+v3.01.pdf>
- Cook, D. A., Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, 119(2), 166.e7-166.e16. Retrieved from <https://doi.org/10.1016/j.amjmed.2005.10.036>
- Dawson, S., & Siemens, G. (2014). Analytics to literacies: The development of a learning analytics framework for multiliteracies assessment. *International Review of Research in Open & Distance Learning*, 15(4), 284–305.
- Dobson, S. (2003). Review of literacy in the new age by Gunther Kress. *Seminar Net: Media, Technology & Lifelong Learning*. Retrieved from www.seminar.net
- Doorey, N. (2014). *Smarter balanced 'tests of the test' successful: Field test provides clear path forward* (research). Smarter Balance Assessment Consortium. Retrieved from https://www.smarterbalanced.org/wp-content/uploads/2015/08/FieldTest_Report.pdf
- Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues & Practice*, 29(2), 24–35. <https://doi.org/10.1111/j.1745-3992.2010.00173.x>
- Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.) London: SAGE Publications Ltd.
- Fitzpatrick, B., & Schulz, H. (2015). Do curriculum outcomes and assessment activities in science encourage higher order thinking? *Canadian Journal of Science, Mathematics*

and Technology Education, 15(2), 136–154. Retrieved from

<https://doi.org/10.1080/14926156.2015.1014074>

Griffin, P., & Care, E. (2014). Test construction. In P. Griffin (Ed.), *Assessment for teaching* (pp. 156–173). Melbourne, VIC: Cambridge University Press.

Griffin, P., & Robertson, P. (2014). Writing assessment rubrics. In P. Griffin (Ed.), *Assessment for teaching* (pp. 107–124). Melbourne, VIC: Cambridge University Press.

IBM Corp. (2012). IBM SPSS Statistics for Windows, Version 21 (Version 21). Armonk, New York: IBM.

Jewitt, C. (2003). Re-thinking assessment: Multimodality, literacy and computer-mediated learning. *Assessment in Education: Principles, Policy & Practice*, 10(1), 83–102. Retrieved from <https://doi.org/10.1080/09695940301698>

Johnson, D., & Kress, G. (2003). Globalisation, literacy and society: Redesigning pedagogy and assessment. *Assessment in Education*, 10(1), 5–14. Retrieved from <https://doi.org/10.1080/09695940301697>

Kalantzis, M., Cope, B., & Harvey, A. (2003). Assessing multiliteracies and the new basics. *Assessment in Education: Principles, Policy & Practice*, 10(1), 15–26. Retrieved from <https://doi.org/10.1080/09695940301692>

Kimber, K., & Wyatt-Smith, C. (2010). Secondary students' online use and creation of knowledge: Refocusing priorities for quality assessment and learning. *Australasian Journal of Educational Technology*, 26(5), 607–625.

Kress, G. (2003). *Literacy in the new media Age*. London: Routledge.

Lemons, P. P., & Lemons, J. D. (2013). Questions for assessing higher-order cognitive skills: It's not just bloom's. *CBE - Life Sciences Education*, 12(1), 47–58.

- Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching*. Boston : Pearson, c2013. Retrieved from <http://ezproxy.uow.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cat03332a&AN=uow.b2007715&site=eds-live>
- Messick, S. (1990). Validity. In R. Linn (Ed.), *Educational Measurement* (3rd ed.), (pp. 13–103). New York: American Council on Education and Macmillin. New London
- Group. (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review*, 60(1), 60–92.
- Rasch, G. (1960/1980). *Foreword and introduction. Probabilistic models for some intelligence and attainment tests..* Chicago: The University of Chicago Press.
- RUMM Laboratory Pty Ltd. (2009). Interpreting RUMM2030: Part I dichotomous data. RUMM Laboratory Pty Ltd.
- Starkey, L. (2011). Evaluating learning in the 21st century: A digital age learning matrix. *Technology, Pedagogy and Education*, 20(1), 19–39. Retrieved from <https://doi.org/10.1080/1475939x.2011.554021>
- ThinkLink Consultancy. (2014). *NAPLAN online 2014 development study cognitive interviews research activity 3: Technically enhanced items (reading)*. Adelaide: ThinkLink.
- Turcsányi-Szabó, M. (2012). Aiming at sustainable innovation in teacher education -- from theory to practice. *Atsinaujinanči Ū Inovacijų Siekimas Mokytojų Rengime - Nuo Teorijos Prie Praktikos.*, 11(1), 115–130.
- Unsworth, L. (2002). Changing dimensions of school literacies. *Australian Journal of Language and Literacy*, 25(1), 62+.

Appendix A:

Example 1 – Question 13, ‘Understanding’, Topic – Health and Nutrition

Section 1 - Question 13

Question

According to the presenter in the video why is high consumption of salt a problem?

Listen

Answer

It can lead to:

- A ☐ Dehydration after a few hours
- B ☐ Heart disease later in life
- C ☐ A salty taste left in your mouth after a few hours
- D ☐ Heart burn later in life

Back

Next

Example 2 – Question 30, ‘Synthesising’, Topic – Social Media

Question 30 - Social Media

Question

Below you have access to a variety of resources (text, audio and video).

Using this information complete the table below by identifying two risks associated with social networking sites and explain how each risk can be reduced.

Listen

Social Networking Safety

Social networking through online media has become very popular with young people, comprising 64% of the total time they spend online. Many young people engage in online social networks as a 'social lifeline' to provide them with a sense of connectedness. With more and more people embracing modern technology and joining social networking sites, it is important to keep children and young people safe as they navigate their way through these online communities.

What are the risks?

Risks associated with the use of modern technology and social networking sites include but are not limited to:

- Cyberbullying - social networking sites are often used to commit cyber bullying behavior
- Loss of privacy - posting of identifying information on social networking sites can lead to harassment and identify theft
- Viruses and spywares - these can spread to the computer when using social networking sites, particularly when they access file-sharing links posted by their online 'friends'
- Copyright law - many material posted online are protected by copyright law which means it is illegal to reprint or share the material without permission
- Plagiarism - occurs when a person claims they wrote or made something created by another person.
- Validity of information - there is an over-abundance of information online and not all of it is factual or unbiased

Audio:

Answer

Risk	How was it overcome?
Risk 1	How was it overcome?
Risk 2	How was it overcome?

[Back](#)[Next](#)

Disclaimer: These items use varying modes which are not accessible via paper

- **Locating**: searching for and identifying information
- **Understanding**: interpreting the nature, meaning and purpose of information.
- **Applying**: using information appropriate for specific purposes
- **Analysing**: separating information into smaller elements
- **Evaluating**: judging and critiquing the quality and accuracy of information
- **Synthesising**: drawing together relevant information into a coherent whole
- **Creating**: developing and generating products
- **Communicating**: selecting the mode of communication to share with an audience.

Figure 1. Model of Multiliteracy skills.

Table 1. Structure of the o-Mlit

Section	Time to complete	Modes used	Response type	Skill assessed	Total number of items
1	30 min	Text; image; audio; text and image; video (no text); video (text, audio and image)	Multiple choice	Locating	12
				Understanding	10
				Applying	3
2	45 min	Text; image; text and image; text, image, audio and video	Short answer	Analysing	5
				Evaluating	3
				Synthesising	2
3	30 min - planning 60 min - producing	Text; image; audio; text and image; video (no text); video (text, audio and image)	Presentation	Creating	4
				Communicating	5

Figure 2. Procedures for o-MLit testing

Table 2. Themes and Codes for the Pilot Studies

Theme	Definition	Pilot Study One		Pilot Study Two - on Revised o-Mlit Assessment	
		Code	Number of Instances	Code	Number of Instances
Test Items	Focuses on how the students interact and understand the items including the source material such as not understanding what was required of them by the item or not using the source material to answer an item.	a – mispronounced words and/or did not know the meaning of a word	6	a – mismatch between item and multiple choice options	5
		b – did not understand what was required of the item	11	b – did not understand what was required of the item	3
		c – focussed on other information in the item	11	c – focussed on other information in the item	1
		d – engagement with source material	10	d – engagement with the source material	6
		e - text in source material (amount, language, formatting)	7	e - text in source material (amount, language, formatting)	2
		f - wording multiple choice options	5	f – wording of multiple choice options	8
				g – item wording	5
Technical Issues	Focuses on the ability to navigate through o-Mlit, and the layout of the items and the source material such as consistency between items, identifying the source material and the inability to open source materials.	a - consistency of layout between items and/or topics	1	a – formatting of the item	5
		b - overall structure (instructions, navigation)	5	b - overall structure (instructions, navigation)	6
		c - clearly identifying source material	3		
		d – technical issues (source not working, incorrect sources)	6		

Table 3. Test/retest reliability for each skill

Skill (Number of items)	Occasion 1 - Occasion 2		Occasion 2 – Occasion 3		Occasion 1 – Occasion 3	
	n	<i>r</i>	n	<i>r</i>	n	<i>r</i>
Locating (12)	272	.60***	270	.65***	271	.53***
Understanding (10)	272	.58***	270	.52***	271	.58***
Applying (3)	272	.08	269	.33***	270	.11
Analysing (5)	271	.50***	263	.50***	265	.55***
Evaluating (3)	271	.27***	263	.31***	265	.22***
Synthesising (2)	271	.37***	263	.47***	265	.36***
Creating (3)	147	.39***	194	.37***	161	.13
Communicating (4)	151	.38***	196	.39***	164	.14

***p<.001

Table 4. Internal Reliability (α)

Skill (number of items)	Occasion 1		Occasion 2		Occasion 3	
	n	α^*	n	α^*	n	α^*
Locating (12)	288	.61*	283	.68*	271	.73*
Understanding (10)	288	.62*	283	.59	271	.60*
Applying (3)	288	.02*	283	.19*	271	.32
Analysing (5)	289	.43*	281	.56*	272	.39*
Evaluating (3)	289	.20*	281	.17*	272	.33*
Synthesising (2)	289	.13	281	.21	272	.19
Creating (3)	187	.47	227	.59	250	.54*
Communicating (4)	190	.75*	229	.78*	250	.73*

*increases to α if item-deleted

Figure 3. Average Skill difficulty

Table 5. Items in order of mean difficulty with fit statistics from each testing occasion

Occasion 1						Occasion 2						Occasion 3					
Question Number (Skill)	Mean Item Difficulty (Logits)	Item Difficulty (Logits)	N	Residual Statistic	Chi-Square Fit Statistic	Probability of fit to the Rasch Model	Item Difficulty (Logits)	N	Residual Fit Statistic	Chi-Square Fit Statistic	Probability of fit to the Rasch Model	Item Difficulty (Logits)	N	Residual Fit Statistic	Chi-Square Fit Statistic	Probability of fit to the Rasch Model	
1a (L)	-2.56	-2.78	275	0.04	4.43	0.35	-2.50	269	-0.24	9.90	0.04	-2.41	258	0.12	7.40	0.12	
11 (U)	-2.27	-2.20	273	-0.83	7.35	0.12	-2.15	265	-0.76	6.91	0.14	-2.47	254	-0.02	3.06	0.55	
2a (L)	-1.83	-1.76	275	0.08	2.16	0.71	-1.79	269	-0.94	15.01	0.00	-1.94	258	-1.47	11.59	0.02	
7 (L)	-1.72	-1.33	269	0.05	1.70	0.79	-1.76	263	-1.12	9.57	0.05	-2.07	255	-1.49	20.63	0.00	
18 (U)	-1.61	-0.91	247	0.96	6.15	0.19	-1.68	260	0.12	11.94	0.02	-2.23	249	-0.81	3.99	0.41	
17 (U)	-1.41	-0.02	260	1.03	6.25	0.18	-1.89	260	0.87	7.07	0.13	-2.33	250	-0.77	12.85	0.01	
4b (L)	-1.27	-1.34	273	-0.65	6.24	0.18	-1.16	266	-1.02	8.77	0.07	-1.30	256	-1.21	3.52	0.47	
4a (L)	-1.19	-1.13	275	-1.55	8.82	0.07	-1.09	269	-1.05	7.50	0.11	-1.37	258	-1.49	9.24	0.06	
32^ (Co)	-1.10	-0.46	253	0.08	2.86	0.58	-2.17	261	-1.15	4.26	0.37	-0.67	254	1.20	2.97	0.56	
1b (L)	-0.70	-0.52	273	-0.48	1.38	0.85	-0.62	267	-1.71	9.77	0.04	-0.94	255	-1.35	10.92	0.03	
13 (U)	-0.55	0.24	257	0.07	6.26	0.18	-0.70	262	-0.78	6.50	0.16	-1.19	253	0.48	6.25	0.18	
35 (Cr)	-0.33	-0.73	178	-1.05	10.56	0.03	0.07	215	-0.95	10.40	0.03	-0.37	237	-1.02	9.92	0.04	
34 (Cr)	-0.21	-0.36	178	-0.01	1.42	0.84	0.23	215	-1.24	6.19	0.19	-0.52	237	-0.04	7.19	0.13	
21 (Ap)	-0.19	0.11	150	0.00	4.65	0.32	-0.28	247	-1.04	9.69	0.05	-0.04	247	-0.72	5.50	0.24	
14* (U)	-0.18	0.46	260	4.21	18.76	0.00	-0.33	259	-0.44	5.91	0.21	-0.67	255	-1.64	14.88	0.00	
3b (L)	-0.17	0.22	272	0.59	4.79	0.31	-0.04	268	0.56	1.63	0.80	-0.48	256	-1.99	14.73	0.01	
5 (L)	-0.11	0.22	275	-0.86	6.77	0.15	-0.11	265	-2.02	7.61	0.11	-0.45	256	-1.57	8.67	0.07	
38 (Co)	-0.11	0.13	178	-0.20	8.61	0.07	0.02	215	-0.22	11.71	0.02	-0.48	237	0.25	1.31	0.86	
8 (L)	-0.11	-0.07	274	-0.59	9.51	0.05	-0.02	268	-0.99	8.36	0.08	-0.26	258	-1.45	7.86	0.10	
3a (L)	-0.06	0.11	275	-1.98	9.18	0.06	0.06	269	-1.48	7.67	0.15	-0.36	269	-2.30	17.81	0.00	
24^ (An)	-0.01	0.14	273	-0.80	5.38	0.25	-0.05	267	0.14	5.18	0.27	-0.13	259	1.15	7.40	0.12	
23# (An)	0.16	0.11	274	0.98	12.57	0.01	0.06	267	0.64	3.17	0.53	0.31	259	1.82	26.32	0.00	
30* (S)	0.20	0.24	247	2.94	2.82	0.59	0.24	267	2.72	15.27	0.00	0.12	259	-2.38	15.30	0.00	
19* (Ap)	0.31	0.70	247	4.26	19.90	0.00	0.13	254	3.44	10.71	0.03	0.09	253	2.19	8.73	0.07	

Question Number (Skill)	Occasion 1						Occasion 2						Occasion 3					
	Mean Item Difficulty (Logits)	Item Difficulty (Logits)	N	Residual Statistic	Fit Chi-Square	Probability of fit to the Rasch Model	Item Difficulty (Logits)	N	Residual Fit Statistic	Chi-Square Fit Statistic	Probability of fit to the Rasch Model	Item Difficulty (Logits)	N	Residual Fit Statistic	Chi-Square Fit Statistic	Probability of fit to the Rasch Model		
33 (Cr)	0.37	0.47	178	-0.08	4.94	0.29	0.41	215	-0.99	6.20	0.18	0.23	237	-0.88	3.46	0.48		
15 (U)	0.37	0.62	264	-0.66	9.91	0.04	0.26	255	-1.56	2.93	0.57	0.25	256	-1.65	8.51	0.07		
10* (U)	0.41	0.77	269	-2.01	11.91	0.02	0.32	262	-3.60	23.87	0.00	0.14	254	-5.19	32.42	0.00		
12 (U)	0.64	0.91	263	0.47	5.14	0.27	0.73	263	0.99	0.33	0.99	0.28	256	-0.33	1.52	0.82		
29 (E)	0.71	0.65	252	0.68	3.63	0.46	0.77	267	0.45	1.83	0.77	0.70	259	0.00	5.76	0.22		
20 (Ap)	0.78	1.68	243	2.07	12.48	0.01	0.46	257	-0.12	1.47	0.83	0.19	251	-0.63	3.5	0.43		
36*^ (Cr)	0.78	0.95	178	1.97	8.72	0.07	0.79	215	1.55	8.56	0.07	0.62	237	2.67	11.74	0.02		
26^ (An)	0.79	0.87	273	0.24	2.51	0.64	0.66	267	0.09	1.48	0.83	0.86	259	-0.81	1.48	0.83		
25 (An)	0.87	0.76	272	1.55	4.43	0.35	0.95	267	0.53	2.18	0.70	0.90	259	0.71	7.92	0.09		
37^ (Co)	0.89	1.01	178	2.06	25.26	0.00	0.78	215	1.46	10.80	0.03	0.87	237	1.14	10.04	0.04		
27*^ (E)	0.90	0.91	272	2.64	12.29	0.02	1.00	267	3.43	23.41	0.00	0.79	259	-0.34	2.86	0.65		
9 (U)	0.90	1.02	273	1.32	8.85	0.06	0.95	267	1.32	22.05	0.00	0.73	256	-1.95	8.76	0.07		
6 (L)	0.94	1.05	270	-0.35	7.36	0.12	0.95	263	-0.92	5.52	0.24	0.82	256	-0.99	2.45	0.65		
22* (An)	0.96	1.22	262	2.76	14.45	0.01	0.99	267	0.52	6.39	0.17	0.66	259	1.07	6.01	0.20		
31# (S)	1.08	1.79	247	2.11	19.97	0.00	0.77	267	1.72	1.22	0.87	0.68	259	2.86	5.24	0.26		
2b (L)	1.29	1.47	273	0.50	5.73	0.22	1.27	260	0.96	5.23	0.26	1.14	256	1.59	4.96	0.29		
16 (U)	1.44	1.92	263	-0.20	11.02	0.03	1.23	261	-0.14	6.68	0.15	1.17	257	2.39	3.93	0.42		
28*^# (E)	2.91	2.99	256	1.11	11.53	0.02	3.01	267	1.23	5.74	0.22	2.72	259	1.97	20.53	0.00		

L = Locating, U = Understanding, Ap = Applying, An = Analysing, E = Evaluating, S = Synthesising, Cr = Creating, Co = Communicating

* This item has been investigated further as it has violated the fit-residual (+/- 2.5) and Chi-square probability (<.05)

^ This item has been investigated because it has violated both the category probability and threshold order

This item has been investigated because there was DIF for topics on all three occasions.

Figure 4. ICCs for Item 19 (a) and 10 (a)

(a) ICC for Item 19 occasion 1

(b) ICC for Item 10 occasion 3

Figure 5. ICCs for Item 14 on occasion 1 (a), 2 (b) and 3 (c)

(a)

(b)

(c)