Development and validation of a concept inventory for introductory-level climate change science

Lorna Jarrett[a], Brian Ferry[b], and George Takacs[a]

Corresponding author: lorna.e.jarrett@gmail.com
[a]Faculty of Engineering, University of Wollongong, NSW 2520, Australia
[b]Faculty of Education, University of Wollongong, NSW 2520, Australia

Keywords: concept inventory**,** climate change, psychometrics**,** validation

International Journal of Innovation in Science and Mathematics Education, *20*(2), 25-41, 2012.

Abstract

This paper follows on from Jarrett, Takacs and Ferry (2011) which reported the first stage in development of a high school level concept inventory (CI) for the science of climate change: the climate change concept inventory (CCCI). In order to develop a reliable and valid instrument, it is necessary to follow appropriate procedures. This paper details the process of CI item development; reports statistical results of initial field trials and outlines how these will be used to further refine the CCCI. Item difficulty, discrimination, and point biserial coefficient were calculated for each item. Cronbach's alpha and test-retest data were used to assess reliability. Results suggest that about half of the items were too difficult for high school students. However, item discrimination and test reliability values were close to acceptable values, which suggests that most students were not simply guessing answers.

Although it was initially designed for use in high schools, a group of undergraduates trialled the CI. Statistical analyses of scores suggest that for this group, the items performed better, and well within acceptable values. Given these favourable results and the fact that introductory-level climate change is increasingly taught at universities, further trials with undergraduates are taking place. It is intended that the final CI will be made available as a formative assessment instrument. The current version is available from the authors on request.

Background and purpose

This research is part of a PhD study which aims to explore high school students' understanding of key concepts underlying the science of climate change. This work builds on a pilot study which concurred with a large body of existing literature that misconceptions about the science of climate change are extremely common (Andersson & Wallin, 2000; Boyes & Stanisstreet, 2001; Hansen, 2010; Koulaidis & Christidou, 1999; Meadows & Wiesenmayer, 1999; Österlind, 2005; Plunkett & Skamp, 1994; Pruneau, Liboiron, Vrain, Gravel, Bourque, & Langis, 2001; Rye, Rubba, & Wiesenmayer, 1997; Schultz, 2009; Shepardson, Niyogi, Choi, & Charusombat, 2009). Several authors have suggested explanations for this, including that students misunderstand key scientific concepts underlying the topic. However, these ideas have not been directly tested.

The PhD study employs multiple data-collection methods to enhance validity. One of these is a concept inventory (CI) which was developed to address seven key conceptual areas underlying the science of climate change. Although a concept inventory had been developed for the mechanism of the greenhouse effect (Keller, 2006), it was not suitable for use in this study for three reasons. These were: lack of comprehensive development and validation

process; the fact that it was developed for undergraduates; and the focus on only one concept related to climate change. Therefore it was necessary to develop an instrument. This paper reports on the development and initial trials of the climate change concept inventory (CCCI). It follows on from Jarrett, Takacs and Ferry (2011) which described stage 2 in the development process as shown in Table 1, ie: the process used to define the conceptual areas to be tested by the CI items. This paper describes stages 3, 4 and 5, ie: development of test items, field trials and data analysis. The results reported in this paper focus on validation through statistical measures of item and whole-test performance. This is an important stage in the study because it establishes the degree to which students' responses to the test can be used to make valid inferences about their conceptual understanding. This is similar to the work of Lindstrøm and Sharma (2010), which describes the development and validation process for a survey designed to measure goal orientation in tertiary physics students: this work was necessary because no suitable instrument already existed.

Concept inventories in science, mathematics and engineering

A concept inventory is a multiple-choice instrument designed to assess students' conceptual understanding of a topic and diagnose misconceptions (Libarkin, 2008). Concept inventories originated in physics education with the Force Concept Inventory (Hestenes, Wells, & Swackhamer, 1992) but have since been developed for many other areas in physics, other sciences, engineering, mathematics, statistics and computing (Anderson, Fisher, & Norman, 2002; Bardar, Prather, & Brecher, 2006; Ding, Chabay, Sherwood, & Beichner, 2006; Gray, Costanzo, Evans, Cornwell, Self, & Lane 2005; Herman, 2011; Libarkin & Anderson, 2006; Lindell & Olsen, 2002; Martin, Mitchell, & Newell, 2004; Pavelich, Jenkins, Birk, Bauer, & Krause, 2004; Rhoads & Roedel, 1999; Richardson, Steif, Morgan, & Dantzler, 2003; Smith, Wood, & Knight, 2008; Stone, 2006; Tongchai, Sharma, Johnston, Arayathanitkul, & Soankwan, 2009; Wuttiprom, Sharma, Johnston, Chitaree, & Soankwan, 2009; Yeo & Zadnik, 2001).

Concept inventories have a number of advantages as data-collection instruments. They can be used to collect data from a large number of participants; they are suitable for exploring participants' ideas about a number of concepts; and because the item distractors are based on known misconceptions they can be used to assess the prevalence of misconceptions. They also have an advantage over open-ended tasks in that students are prompted to choose an option; this is important because the pilot study on which this research was based showed that participants often volunteered very little information without prompting, but when prompted were able to express their ideas in more detail.

Literature on concept inventory development and validation

Validity and reliability

An important difference between a concept inventory and any other multiple-choice test is the focus on validity and reliability, which means that students' responses to the CI can confidently be assumed to reflect their knowledge of the topic. The development and validation process for the climate change concept inventory was based on the advice of a number of authors.

Most authors agree that validity derives from a rigorous development process, so validity must be considered from the start of the CI development process. For example, an early stage in development is concept selection. Gray et al. (2005) asserted that using the Delphi process

for concept selection should, if carried out correctly, contribute to validity. The Delphi process is a method of structuring group discussion through multiple iterations of a survey, with feedback provided to participants (Jarrett et al., 2011).

According to Libarkin (2008), multiple forms of validity must be considered at all stages of CI development to ensure the resulting instrument is effective. Richardson et al. (2003) described a CI development and validation process following the development of an unsuccessful CI which had been generated using a less rigorous process. Steif and Dantzler (2005) state that validity is more difficult to establish than reliability and recommends assessing multiple forms of validity. The most common form of validity mentioned in CI development literature is content validity, defined as adequate coverage of the content domain (DeVellis, 2003; Gray et al., 2005; Steif & Dantzler, 2005). Approaches to content validity include use of known misconceptions in item distractors (Steif and Dantzler, 2005); expert review of items (Bardar et al., 2006; Ding et al., 2006); use of Delphi method and multiple focus groups in the development process (Gray et al., 2005) and use of established instrument design processes including item-writing guidelines (Libarkin, 2008).

Miller (1995) defined test score reliability as the ratio of true variance, due to differences between people sitting the test, to total variance. In other words, reliability measures illustrate how much of the variation in test results is due to measurement error (Steif & Dantzler, 2005). Miller described three methods for estimating reliability: test-retest, alternative-forms and internal-consistency. The author pointed out that the less homogenous a test is, the lower the estimate of reliability an internal-consistency measure will give. Concept inventories typically measure understanding of a number of concepts so a good concept inventory could not be expected to be homogenous, as good understanding of one of the concepts may not imply good understanding of others. Therefore measures of internal consistency can 'badly under-estimate reliability' (Miller, 1995; p. 270) of non-homogenous tests such as concept inventories. Similarly, Gray et al. (2005) described internal-consistency measures as conservative estimates of reliability for test results.

Stages in concept inventory development and validation

Table 1 shows the main stages in CI development, methods used by CI authors in the literature, and methods used in our research.

Development and validation of the climate change concept inventory

Stage 1: Identifying the purpose of the concept inventory

The purpose of developing the concept inventory was to investigate students' understanding of the key scientific concepts underlying the most basic scientifically-acceptable explanation of the science of climate change.

Stage 2: Choosing concepts

Content validity, as described in Section 2.2.1, requires adequate coverage of relevant concepts. The list of concepts to be covered can come from a variety of sources e.g. consultation with teaching staff and students, literature on the topic, or theory. Delphi studies have been used to structure consultation for this stage of CI development (Danielson, 2005; Gray et al., 2005; Herman, Loui, & Zilles, 2010; Rowe & Smaill, 2007; Stone et al., 2004; Streveler, Olds, Miller, & Nelson, 2003). The list of conceptual statements for the climate change concept inventory was derived from a synthesis of two ranked lists of concept

statements: one from a Delphi study involving discipline experts and the other from a literature review of research into students' understanding of the science of climate change, focused on studies involving participants of similar age to the high school students in this study. This process was reported in Jarrett et al. (2011).

Table 1: Stages in concept inventory development and validation

Stage in CI development	*Examples of methods adopted in the literature*	*Methods adopted for climate change concept inventory*
1. Identify purpose	Identify the primary purpose for which test scores will be used (Bardar et al., 2002).	PhD research questions: what do high school students know about the concepts underlying climate change?
2. Choosing the list of concepts	Expert opinion and review of course texts (Libarkin and Anderson, 2006). Concepts commonly taught in majority of courses (Bardar et al., 2002). Expert opinion of concepts important for establishing validity: Delphi suggested (Richardson, 2004). Based on an existing open-response instrument (Tongchai et al., 2009)	Conceptual statements based on Delphi study and review of literature.
3. Initial item development	Distractors based on prior research into students' ideas, teaching experience and student interviews. Draft questions reviewed by experts. Small scale trials followed by interviews. Revision based on statistics and feedback (Bardar et al., 2002). Distractors based on responses to open-ended questionnaire and open-ended interview, with interview protocol based on list of topics. (Libarkin and Anderson, 2006). Two-stage process starting with open-ended stems. Distractors based on student interviews, focus-groups and answers to open-ended questions (Richardson, 2004; Gray et al., 2005).	Literature study for known misconceptions on concepts identified in stage 1. Open-ended questions written based on concepts identified in stage 1, trialled with four focus groups to determine how questions are interpreted and identify misconceptions. Development of items with distractors based on misconceptions from literature review and focus groups. Draft CI items sent to Delphi participants for review and revised as necessary. Second round of focus groups using 'think-aloud' protocol with revised draft CI items.
4. Initial field trials	Minimum acceptable sample size is 5–10 times as many subjects as test items (Nunnally, 1967).	229 students in years 9 and 10 68 undergraduates, for a 27 item test.
5. Data analysis	See Table 4.	See Table 5.
6. Revision	Think-aloud interviews (Libarkin and Anderson, 2006). Eliminate items that do not meet pre-established criteria (Bardar et al., 2002). Improve readability, validity, reliability and fairness (Richardson, 2004). Experts review items and suggested revisions. Post-tests using revised questions (Libarkin and Anderson, 2006).	Four focus groups with high-school students for validation of CI responses. Revision and development of beta version.

Stage 3: Initial item development

Development of items for the climate change concept inventory was informed by item-writing guidelines summarised in table 2. These were based on the following sources: Haladyna, Downing, & Rodriguez's (2002) who synthesised a list of 31 guidelines for writing multiple-choice test items, from a review of 27 textbooks; Frey, Petersen, Edwards, Pedrotti, & Peyton (2005), who compiled a list of rules from assessment textbooks; and advice on CI item development from Libarkin (2008) and Libarkin and Anderson (2006).

Table 2: Summary of item-writing guidelines

1	Structure the stem as a question if possible (Libarkin, 2008; Haladyna et al., 2002).
2	Language should be as simple as possible and appropriate to the target population. Minimise the amount of reading in each item (Libarkin, 2008; Bardar et al., 2006).
3	Stems should be unambiguous (Libarkin, 2008), should clearly state the problem and directions. Include the central idea in the stem rather than in the options.
4	Ensure that all distractors are plausible (Libarkin, 2008): use known misconceptions (Bardar et al., 2006).
5	Avoid 'type K' formats and 'all of the above' as an option. Use 'none of the above' sparingly (Haladyna et al., 2002) or not at all (Libarkin, 2008; Frey at al., 2005).
6	Avoid giving hints to correct answer by keeping options homogenous in length, content and grammatical structure; the correct option should not be the longest one (Frey et al., 2005).
7	Avoid trick and opinion-based questions or trivial content (Haladyna et al., 2002)
8	Avoid negatives e.g. NOT or EXCEPT (or capitalise if they must be used). Avoid absolutes such as 'always, never,'etc. (Libarkin, 2008) and vague frequency terms e.g. 'often', 'usually' (Frey et al., 2005).
9	Avoid complexity in responses e.g. 'X and Y but not Z' (Libarkin, 2008).
10	Avoid the same, or a very similar word appearing in stem and correct option (Haladyna et al., 2002).
11	Vary the location of the right answer, place choices in logical or numerical order (Haladyna et al., 2002).
12	Options should be logically independent of each other e.g. there should be no overlap in numerical ranges (Haladyna et al., 2002).
13	Use novel material to test higher-level learning. Paraphrase textbook language to avoid simple recall (Haladyna et al., 2002).
14	Avoid over-specific and over-general content (Haladyna et al., 2002).
15	Three choices are adequate (Haladyna et al., 2002).
16	Options should not have repetitive wording (Frey et al., 2005).
17	Questions of the same format should be together (Frey et al., 2005).
18	Answer options should be available more than once (Frey et al., 2005).
19	Each item should reflect specific content and a single specific mental behaviour (Haladyna et al., 2002).
20	Format options vertically (Haladyna et al., 2002).

Open-ended questions were written to cover the conceptual statements developed in Stage 2. Audio-recorded focus groups, using the open-ended questions, were carried out with four groups of students. Another literature review was carried out to identify known misconceptions about the concepts described in the conceptual statements.

Audio recordings were transcribed and analysed thematically. This involved grouping responses from different focus groups to specific questions together, coding statements using a tool described in Diment (2010); then aggregating phenomenographically equivalent

statements to give a list of conceptual statements (correct and incorrect) for each conceptual area (Stephanou, 2007). These lists, along with misconceptions identified from the literature review, formed the basis of the distractors for items.

A total of 58 items were written in order to enable weeding out of weaker ones. The researchers reviewed these items four times, resulting in the deletion of some items, the revision of others and the creation of several more, to give a final pool of 44 items.

Three conceptual areas were not covered. These were: *the difference between weather and climate*, *past climate* and *radiative forcing capacity*. No items were written on *the difference between weather and climate* because no misconceptions about this conceptual area were observed during focus groups. None were written for *past climate* because all participants had heard of ice ages but no other climactic periods were described ie: responses were homogenous so plausible distractors could not be written. However, it may be possible to devise items asking e.g. whether variation was only caused by human activity. It may be advisable to address this when revising the concept inventory after analysis of field trial data; however it is also important that the test is not excessively long: most concept inventories are designed to be completed within 30 minutes (Richardson, 2004). No items were written for *radiative forcing capacity* because the necessary level of understanding of the interaction between greenhouse gases and infra-red radiation was not observed during any of the focus groups. Again, it may be possible to devise items addressing this concept in subsequent versions.

The 44 items in the final pool were sent to the Delphi study participants for comments and feedback, and then taken to two focus groups. Focus group participants were asked to verbally explain their reasons for choice of response; to flag any words they didn't understand and ask the researcher for explanation if any questions weren't clear. These consultations resulted in further refinement of items and helped in rejecting problematic items.

Following analysis of focus-group data, the researchers determined which items to retain. The aim was to retain 25-30 items. Twenty-seven items were retained: these included at least one question per conceptual area. Not all the questions rejected were considered problematic: the need to keep the total number of items under 30 meant that where several items addressed the same concept, it was necessary to reject some based on personal judgment. Items were then arranged in a logical order. To assist with this, the items were reviewed to determine what information was contained in item stems, and what information was asked for. Items supplying information were placed after questions asking for same information in order to minimise the risk of questions ‘tipping-off’ students. Table 3 shows the conceptual statements from Stage 2 and the corresponding CI item numbers.

Stage 4: Trials

229 high school students in six schools participated in field trials: these included public and private; academically selective and non-selective; suburban, semi-rural and rural schools. Participants comprised eleven whole-class groups and completed the CI during normal class time, taking around 25 minutes to complete the task. All data-collection was guided and supervised by the senior author. The importance of students providing their own answers, whether correct or incorrect, was emphasised. Participants were particularly asked not to choose answers at random, but to leave blank any items where they could not at least make an educated guess.

Table 3: Conceptual statements and corresponding item numbers in the first draft of the climate change concept inventory

Conceptual Statement – broken down into individual concepts	*Items*
1. Carbon cycle and fossil fuels:	
There is a fixed amount of carbon on Earth;	3,22
it is cycled among the atmosphere, biosphere, soils, ocean and rocks.	1
There are both natural and human-induced sources and sinks of greenhouse gases.	2,8,12
Fossil fuels contain carbon that was part of living things millions of years ago.	
The process of burial took this carbon out of the atmosphere-ocean-biosphere cycle.	13
Burning fossil fuels returns this carbon to the cycle.	5
2. Electromagnetic spectrum: There is Infra Red (IR) and Ultra Violet (UV) radiation beyond the visible spectrum: these are all related forms of electromagnetic energy.	
The Sun emits mostly visible radiation and the Earth emits mostly IR.	6 14,24
3. Interactions between greenhouse (GH) gases and electromagnetic radiation:	
Most of the gases that make up the atmosphere are transparent to visible light.	15
Non-GH gases are transparent to IR	16
but GH gases absorb IR: this is the cause of the greenhouse effect.	21
GH gases allow the Sun's visible light in	
but absorb IR emitted by Earth. This is re-emitted in all directions - down as well as up.	11,18 23
Natural climate variability in the past and relationship to CO_2 levels: The climate has been different in the past. Prehistoric climate changes correlate with changes in CO_2 levels, providing evidence for the link between CO_2 levels and global temperatures.	NONE
Difference between weather and climate: Weather is short-term, day-to-day climactic conditions while climate is the longer term average conditions.	NONE
4. Proportions of greenhouse and non-greenhouse gases in the atmosphere: Over 96% of the atmosphere consists of non-greenhouse gases.	9
The atmosphere also contains small amounts of CO_2, CH_4, O_3, N_2O and H_2O and CFCs- all of which are greenhouse gases.	10,19
Water vapour is a variable component of the atmosphere and is the most abundant greenhouse gas.	17
GH gases are not in a distinct atmospheric layer.	
Radiative forcing capacity: Some greenhouse gases have more radiative forcing capacity than others.	NONE
5. Feedback: changing one parameter can have an effect on another parameter which causes a changes in the original parameter. Feedbacks can be negative (ie: tends to return the parameter to its original value)	25
or positive (ie: tends to drive the parameter further away from its original value) e.g. increasing CO_2 raises surface temps causing more water to vaporise, which further raises temperatures.	26,27
6. Equilibrium of energy: there is a balance of energy into and out of the Earth / atmosphere system. A net flow of energy into or out of the Earth / atmosphere system leads to temperature change over time.	7,20
7. Conservation of energy: Energy can change from one form into another but the total amount of all forms of energy remains constant.	4

It must be acknowledged that the sample of participating students is biased in favour of high academic achievement, with six of the eleven participating classes being selective classes. Participating classes were selected by teaching staff at participating schools who may have chosen selective classes because they were considered more likely to participate actively and

benefit from taking part in the research. The inconvenience of participating in research should be balanced by direct benefit to participants; and many students, particularly in focus groups, reported that they had learned something from the activities. Therefore, we consider this choice of participants to be justified.

The CI was also trialled by 68 undergraduates enrolled in an introductory-level unit of study on climate change. Such units of study are increasingly being offered at universities and we wished to assess whether the climate change concept inventory might be suitable for students at this level. Similarly, Hestenes (1992) trialled the Force Concept Inventory with both undergraduates and high school students. The undergraduate participants in this study completed the climate change concept inventory at the end of their unit of study, during normal tutorial time and also took around 25 minutes to complete the task.

Table 4: statistical measures used to establish CI performance in the literature

Name of measure	*Authors and notes on use*	*Recommended values*
Item difficulty (P)	There should be a range of difficulties (Bardar et al., 2006)	0.3-0.9 (Ding et al., 2006); optimum 0.5 0.2-0.8 (Bardar et al, 2006) 0.4-0.6 (Richardson, 2004) <=0.7 (Smith et al., 2008) 0.3-0.7 (Anderson et al. 2002); optimum 0.63 Average 0.5 (Gronlund, 1993; cited in Anderson et al., 2002)
Item discrimination ($D_{25,}$ D_{333} or D_{50})	Assesses how well items distinguish between strong and weak students (Ding et al., 2006)	Must not be –ve (Ding et al., 2006) Good if D>=0.3 (Doran, 1980; cited in Ding et al., 2006; Steif and Dantzler, 2005; Lindell and Olsen, 2002) D_{ave} >=0.3
Point biserial coefficient (r_{pbs})	Consistency of individual items with the test as a whole (Ding et al., 2006)	Negative value indicates defective item; average r_{pbs}>=0.2; few items should have r_{pbs}<0.2 (Ding et al., 2006) All items should be >=0.2 (Kline 1986; cited in Smith et al., 2008) 0.3-0.7 (Allen and Yen, 1979; cited in Bardar et al. 2006; Kaplan and Saccuzzo,1997; cited in Anderson et al., 2002) Minimum 2 SD above 0.00 (Allen and Yen 1979; cited in Bardar et al., 2006)
Cronbach's alpha (whole-test)	Measure of internal consistency; estimate of reliability (Steif and Dantzler, 2005; Bardar et al., 2006 ; Gray et al., 2005; Lindell and Olsen, 2002)	>=0.7 (Litwin 1995; cited in Bardar et al. 2006; Nunally, 1978; cited in Gray et al., 2005)
Kuder-Richardson reliability index (whole-test)	Measure of reliability (Ding et al., 2006); Nottis, Prince, & Vigeant, 2010) Equivalent to Cronbach's alpha for dichotomous data (DeVellis, 2003; Gray et al., 2005)	>=0.7 for groups and >=0.8 for individuals >=0.6 (Gronlund 1993; cited in Anderson et al., 2002)
Test-retest (whole-test)	Measure of reliability (Gray et al., 2005)	Coefficient of stability >0.8 (Smith et al., 2008)

Stage 5: Statistical evaluation

Quantitative analyses are carried out to determine how well concept inventories perform as a measure of students' understanding. These usually include statistical determination of test reliability; item difficulty and discrimination. Statistical measures either apply to single items or to the test as a whole. For single-item measures, the average value for all items (ie: for the test as a whole) is usually also given. Table 4 summarises statistical measures commonly used in the evaluation of concept inventories. Table 5 shows which of these statistical measures were calculated for the climate change concept inventory, along with the widest recommended ranges from literature.

Table 5: statistical measures used in this study

Name of measure	*Recommended values*
Item difficulty (P) = fraction of correct responses	0.2-0.9; optimum 0.5
Item discrimination using top 25% and bottom 25% of scores (D_{25}). Less likely than D_{50} to underestimate item discrimination but involves discarding half of the dataset	>=0.3
Item discrimination using top 50% and bottom 50% of scores (D_{50}). More likely than D_{25} to underestimate item discrimination but includes entire dataset	>=0.3
Point biserial coefficient (r_{pbs}); and associated p value	>=0.2

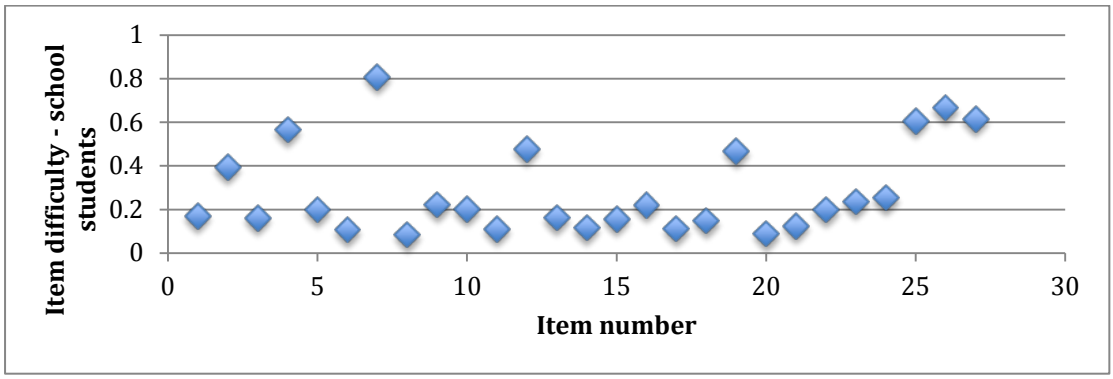
D_{25} and D_{50} values were calculated on Excel and a subset of values from each group hand-checked. r_{pbs} values were calculated using JMP.

Results

High school students

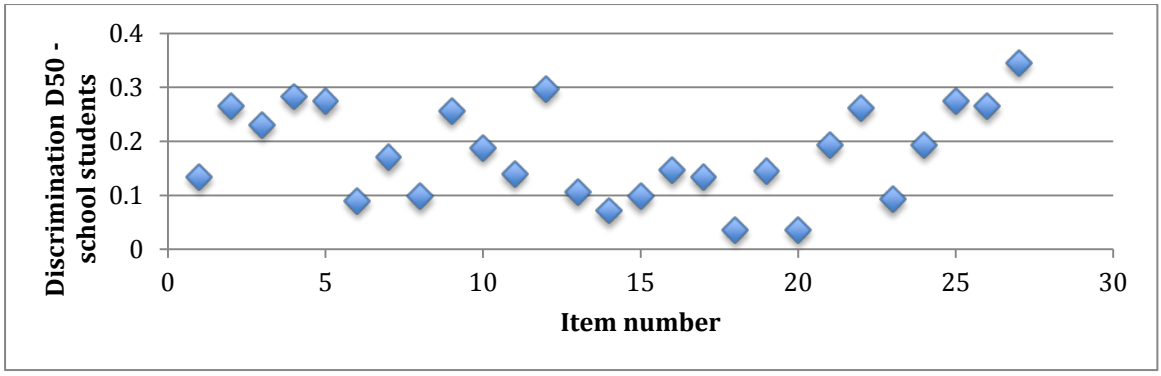
Item difficulty, discrimination and point biserial coefficient

The following four graphs, Figures 1-4, show the results of the statistical evaluation for individual items. Item difficulty, discrimination (D_{25} and D_{50}) and point biserial coefficient are plotted against item number for each of the 27 items in the test. The average value of the corresponding measure is given below each graph, along with the recommended range. These graphs give an indication of which items performed adequately.



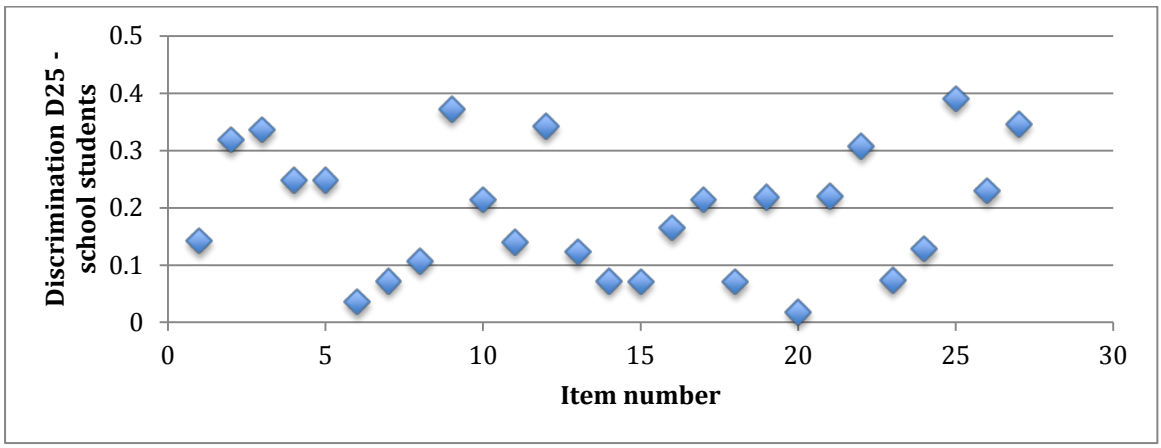
Average item difficulty = 0.28 (cf: acceptable range 0.2-0.9; optimum 0.5)

Figure 1: Item difficulty for high school students



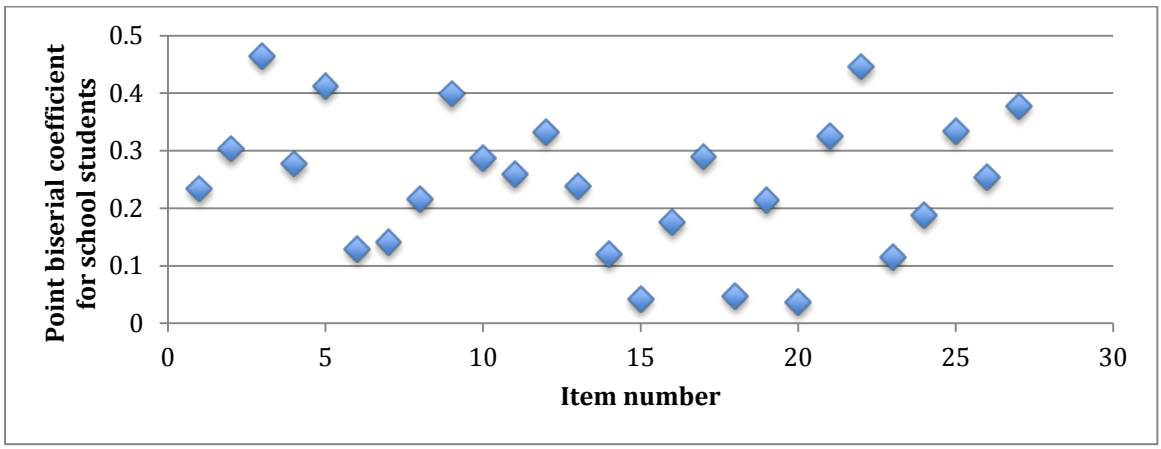
Average D_{50} value = 0.18 (cf: mimimum acceptable 0.3)

Figure 2: Item discrimination D_{50} for high school students



Average D25 value = 0.19 (cf: mimimum acceptable 0.3)

Figure 23: Item discrimination D_{25} for high school students



Average r_{pbs} = 0.247 (cf: minimum acceptable 0.2)

Figure 4: point biserial coefficient for high school students

Measures of reliability

Cronbach's alpha for high school students = 0.6250. This is less than the minimum acceptable value of 0.7 cited by most CI researchers. However Grolund (1993; cited in Anderson et al., 2002) gave 0.6 as a minimum acceptable value for Kuder-Richardson 20 which is equivalent to Cronbach's alpha for dichotomous data. Also, Miller (1995) explained that for CIs which tend not to be homogenous, tests of internal consistency can seriously underestimate reliability. Therefore the value of Cronbach's alpha for high school participant trials does not necessarily indicate that the CI is not sufficiently reliable. Lindell and Olsen's (2002) alpha values were 0.55 for pre-test and 0.75 for post-test. The authors suggested that the low alpha value for the pre-test may be due to students guessing answers to questions where they had no ideas (correct or incorrect) about the concepts. This may possibly be the case for our high school participants, as much of the material covered in the CI is not explicitly taught in schools. Participants were asked to leave questions blank in preference to guessing an answer but they may have made "educated" guesses.

We collected re-test data from 34 students in order to provide an additional reliability measure. It is not sufficient to correlate students' total scores for test and retest: the median score for test data was 7 out of 27 and a student could easily score seven both times while choosing completely different answers. Therefore, for each item we looked at whether each student chose the same response both times.

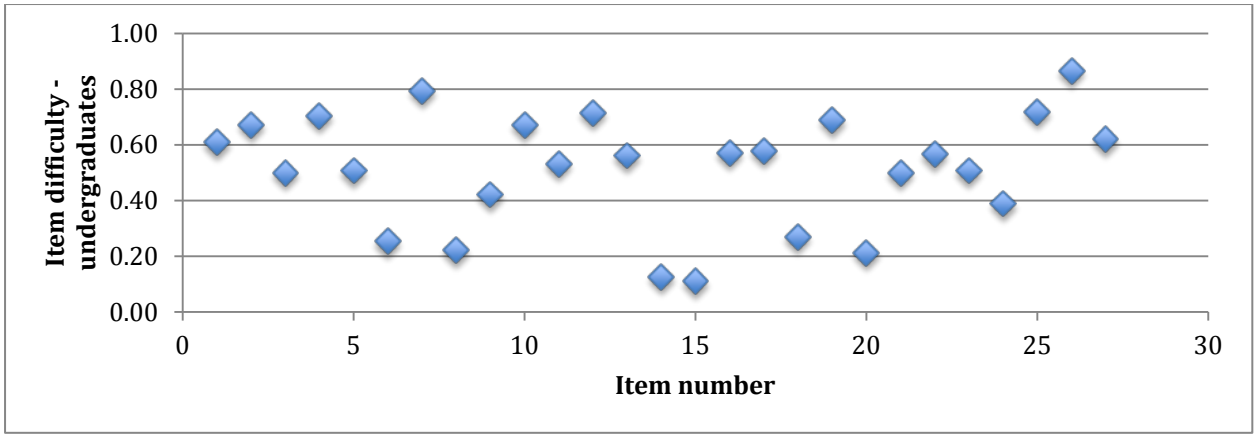
The probability that a student would chose the same item each time on a three item scale if they were answering randomly (i.e. guessing) each time would be .33 and for a four item scale .25. Taking this as the probability of success (i.e. of guessing the same answer twice in two trials) we can use the binomial distribution to calculate the 95% confidence interval of the proportion. If the proportion of students who gave the same response (whether correct or incorrect) twice exceeded the upper bound of the 95% confidence interval then we can conclude that the test-retest group are not guessing for each item where this condition is met. Three of the items had values within the 95% confidence interval ie: for these items the test-retest similarity was not significantly better than if responses had been chosen randomly.

To give an estimate for the whole test we took the average proportion of responses which were the same for test-retest for each item. We then calculated the proportion that would be expected if participants were guessing, using the average number of options for each test. The average number of options for the concept inventory was 3.26, so the reciprocal of 3.26, ie: 0.307, is the proportion of identical responses that would be expected if responses were chosen at random. This gives a confidence interval for the binomial proportion for 3.26 options. If participants were answering at random, the mean probability of choosing the same option both times is 0.307. Therefore the 95% CI of the binomial proportion is 0.1518 to 0.4618. The observed mean proportion for the climate change concept inventory is: 0.562, which lies outside the 95% confidence interval.

Undergraduates

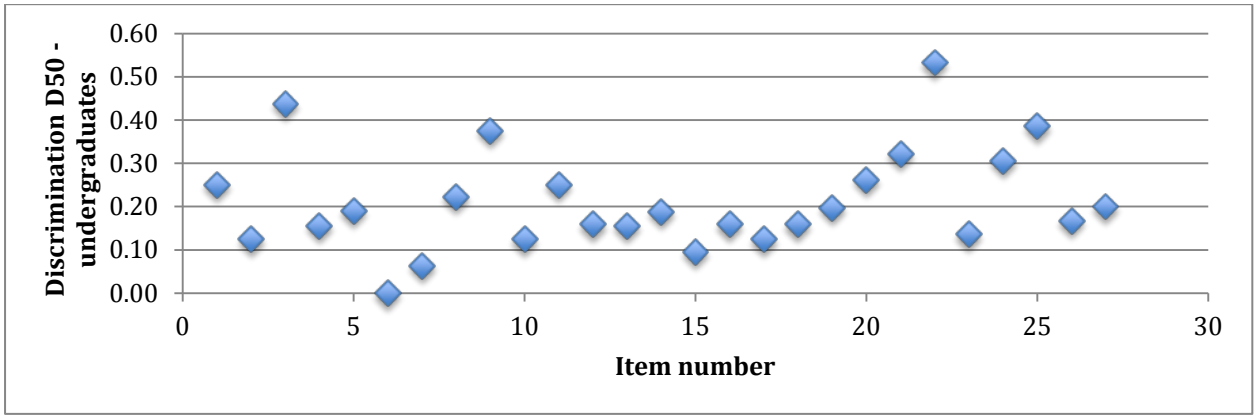
Item difficulty, discrimination and point biserial coefficient

As for the high-school students, the values of item difficulty, discrimination and point biserial coefficient are plotted for each of the 27 items in the test, along with average values in the following four graphs.



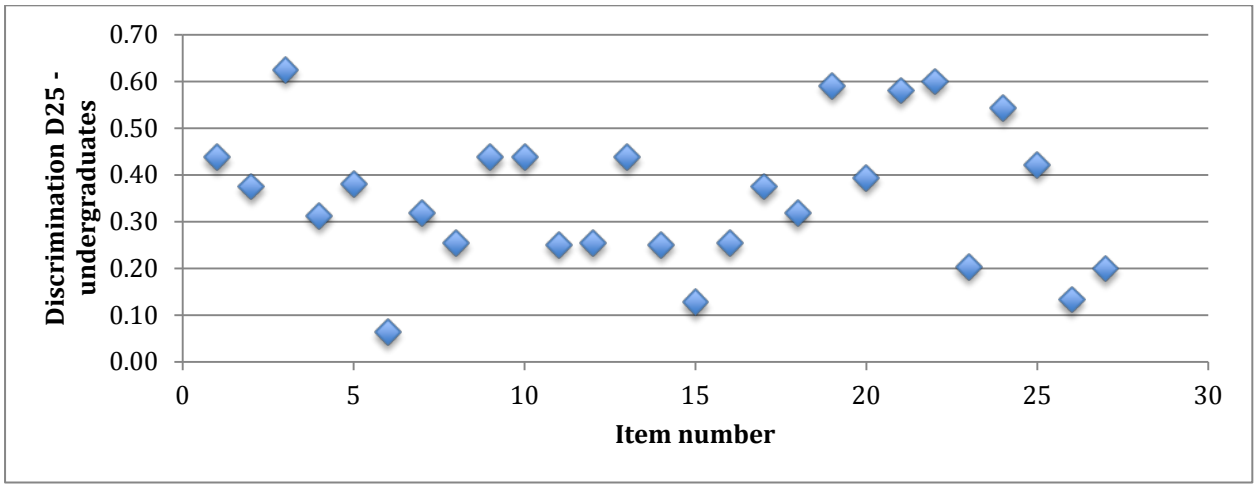
Average item difficulty = 0.51 (cf: acceptable range 0.2-0.9; optimum 0.5)

Figure 5: item difficulty for undergraduates



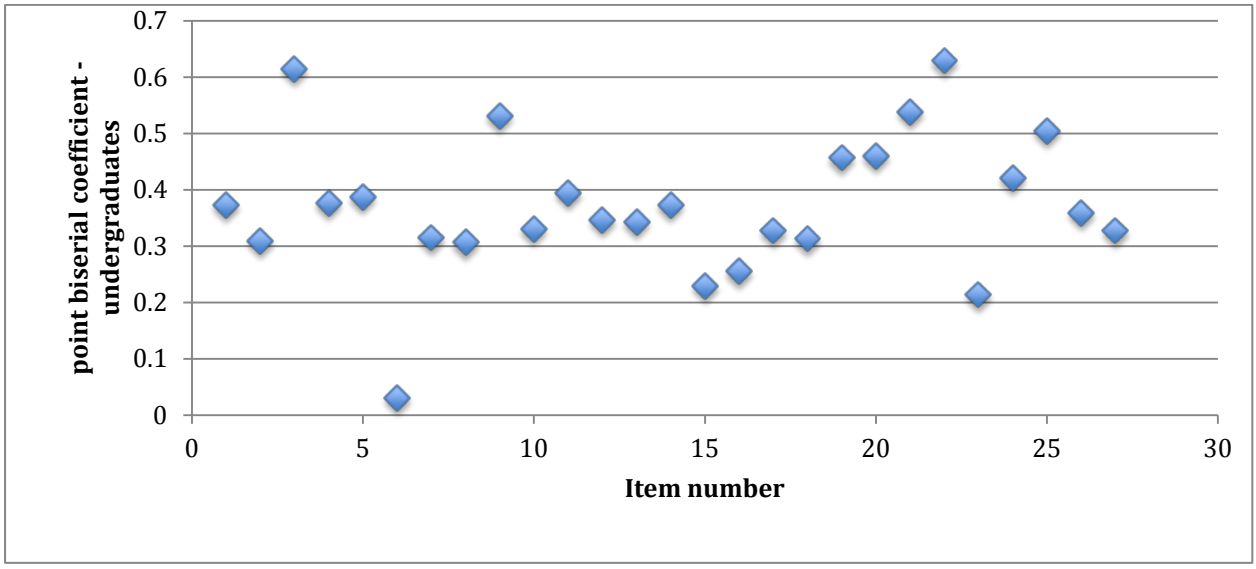
Average D_{50} value = 0.21 (cf: mimimum acceptable 0.3)

Figure 6: item discrimination D_{50} for undergraduates



Average D25 value = 0.35 (cf: mimimum acceptable 0.3)

Figure 7: item discrimination D_{25} for undergraduates



Average rpbs = 0.37 (cf: minimum acceptable 0.2)

Figure 8: point biserial coefficient for undergraduates

Measures of reliability

Cronbach's alpha for undergraduates = 0.7656. This is above the minimum value of 0.7 recommended in literature, and higher than that for the high school group.
It was not possible to collect test-retest data for the undergraduate group. These students were studying the topic at the time so it would have been almost impossible to collect test-retest data without the students having learned something about the topic in the meantime.

Discussion

High school students
Twelve items have difficulty below 0.2 for the high school students: this means that almost half the test items are more difficult than literature recommends. However, the concepts covered were derived from a process (Delphi study and literature review) designed to identify concepts important to a basic understanding of the topic. Therefore the low scores on these items may indicate an unacceptably high rate of misconceptions about these concepts rather than badly-designed questions. Only one item, at 0.81 is above Bardar et al.'s (2006) suggested maximum of 0.8, and none are above Ding et al.'s (2006) suggested maximum of 0.9. Therefore nome of the items are excessively easy.

Although average values for D_{50} and D_{25} for high school students were below the recommended values, the average point biserial coefficient was higher than the discrimination values, and well above the minimum acceptable value. This suggests that the CI as a whole performed reasonably well in discriminating between students with poor and good understanding of the topic. Also, none of the discrimination values or point biserial coefficients were negative, ie: more likely to be answered correctly by someone with a overall poor understanding of the topic than someone with a good understanding. This suggests that none of the items is seriously defective.

When item performance is compared with conceptual area (see Table 3), some patterns emerge. Items 25-27, on feedback, all performed well, as did 9,10,17 and 19, which addressed proportions of greenhouse and non-greenhouse gases in the atmosphere. Items 6,14 and 24, on the electromagnetic spectrum, performed poorly, as did 7 and 20 on equilibrium of energy. Half of the items for interactions between greenhouse gases and electromagnetic radiation (15,18 and 23) performed poorly while the other half (11, 16 and 21) were better. The carbon cycle and fossil fuels generally performed well, with 2,3,5,12 and 22 performing well, and 13 reasonably well; however 8 performed poorly.

Reliability for high school students was on the boundary of acceptability. Although the Chronbach's alpha was below the limit accepted by most researchers, the test-retest data suggested that there was less than 5% probability of the test-retest results being due to chance.

Undergraduates

All statistical measures of test performance were better with the undergraduate group. Although this group was only a third as large as the high school group, it is still large enough for reasonably small effects to be detectable.

Only two items (14 and 15) had difficulty rating outside the recommended range and overall, only three items did not perform well. These were 6, 15 and 23. Item 6 performed poorly (discrimination and point biserial values unacceptably low), 15 less poorly (difficulty and discrimination outside recommended range) and 23 had borderline acceptable performance (discrimination outside acceptable range but point biserial acceptable).

Cronbach's alpha was above the minimum value of 0.7 and as Cronbach's alpha is considered a conservative estimate of reliability for CIs, this suggests that for the undergraduate group, the test was reliable. This result also lends credibility to the idea that the low Cronbach's alpha result for high-school participants was due to their lack of familiarity with the concepts, because the undergraduates, having just completed a unit of study on climate change, would have been more familiar with most of the concepts.

Conclusions

Statistical evaluation of the draft CCCI with undergraduate students shows that most items perform adequately and that the test is reliable. However, most of the items in the draft CCCI appear to be too difficult for high school students, despite the process used to ensure that the test covered appropriate content. This suggests that the high school students have not acquired an understanding of the basic principles needed in order to understand the science of climate change. This idea will be explored in future research.

Several items performed poorly on statistical measures with high school students, compared with results for undergraduates. This may result from the high school participants having tentative mental models of the concepts being tested. There is evidence for this in the post-trial focus-group sessions. For example, participants frequently contradicted themselves, and when asked to explain the relationship between two ideas they had articulated or to follow through a line of reasoning, often stated that they were confused. This lack of concrete mental models may have led to them choosing inconsistent responses to items on related concepts. This idea may be further investigated by trialling the items on participants who are more likely to be familiar with the concepts students e.g. physics undergraduates, to see whether item performance is better.

Future research

The next stage in this study involves examination of high school students' item choices to make inferences about their conceptual understanding, and comparing these to data from post-trial focus groups. This will provide further validation for the concept inventory, as well as allowing participants' reasoning to be examined in more depth and exploring how common misconceptions, such as overestimation of the proportion of greenhouse gases in the atmosphere, or conflation with ozone depletion, may develop.

Another possible line of investigation involves comparing high school students' responses to undergraduates' to determine the effect of learning experiences in overcoming misconceptions.

As described above, some items performed poorly for both groups and it was suggested that both groups may have tentative conceptual models for the corresponding concepts. This explanation can be tested by trialing the concept inventory with participants who can be expected to be more familiar with the concepts, e.g. physics majors. Feedback will also be sought from academics involved in teaching these conceptual areas.

Further trials of the first draft version of the climate change concept inventory are currently taking place with groups of undergraduates: this will provide data to enable refinement of the instruments and trials of a beta version. This may involve writing items for the three conceptual areas that were not covered in the first draft.

Acknowledgements

The authors acknowledge the help, support and contribution of the participating high school students, their classroom teachers and executive staff; the undergraduate students, their tutors and course-coordinators; the University of Wollongong statistical consultant and Kieren Diment for statistics advice; and the Delphi study participants who reviewed draft CI items. This research was supported by an Australian Postgraduate Award.

References

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching*, *39*(10), 952–978. doi:10.1002/tea.10053

Andersson, B., & Wallin, A. (2000). Students' understanding of the greenhouse effect, the societal consequences of reducing CO 2 emissions and the problem of ozone layer depletion. *Journal of research in science teaching*, *37*(10), 1096–1111.

Bardar, E., Prather, E., & Brecher, K. (2006). Development and Validation of the Light and Spectroscopy Concept Inventory. *Astonomy Education Review*, *5*(2).

Boyes, E., & Stanisstreet, M. (2001). Plus ca change, plus c'est la meme chose? School Students' Ideas about the "Greenhouse Effect" a Decade On. *Canadian Journal of Environmental Education (CJEE)*, *6*(1-2001).

Danielson, S. (2005). Developing statics knowledge inventories. *Frontiers in Education, 2004. FIE 2004. 34th Annual* (p. F3G–15).

DeVellis, R. F. (2003). *Scale Development: Theory and Applications* (2nd ed.). SAGE.

Diment, K. (2010). A prototype open source toolkit for comprehensive, flexible and extendable computer assisted qualitative data analysis. Presented at the 5th International Conference on Qualitative Research in IT & IT in Qualitative Research (QualIT2010), Brisbane: Queensland University of Technology.

Ding, L., Chabay, R., Sherwood, B., & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physical Review Special Topics - Physics Education Research*, *2*(1), 010105. doi:10.1103/PhysRevSTPER.2.010105

Gray, G. L., Costanzo, F., Evans, D., Cornwell, P., Self, B., & Lane, J. L. (2005). The dynamics concept inventory assessment test: A progress report and some results. *American Society of Engineering Education, Annual Conference* (pp. 4819–4833).

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education*, *21*(4), 357–364.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, *15*(3), 309. doi:10.1207/S15324818AME1503_5

Hansen, P. J. . (2010). Knowledge about the Greenhouse Effect and the Effects of the Ozone Layer among Norwegian Pupils Finishing Compulsory Education in 1989, 1993, and 2005—What Now? *International Journal of Science Education*, *32*(3), 397–419.

Herman, G. L., Loui, M. C., & Zilles, C. (2010). Creating the digital logic concept inventory. *Proceedings of the 41st ACM technical symposium on Computer science education* (pp. 102–106).

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The physics teacher*, *30*(3), 141–158.

Jarrett, L. E., Takacs, G., & Ferry, B. (2011). What scientific concepts are required to understand climate change? *Proceedings of the Australian Conference on Science and Mathematics Education* (pp. 89–94). Presented at the Australian Conference on Science and Mathematics Education, Melbourne.

Keller, J. M. (2006). *Part I. Development of a concept inventory addressing students' beliefs and reasoning difficulties regarding the greenhouse effect, Part II. Distribution of chlorine measured by the Mars Odyssey Gamma Ray Spectrometer*. The University of Arizona, United States -- Arizona.

Koulaidis, V., & Christidou, V. (1999). Models of students' thinking concerning the greenhouse effect and teaching implications. *Science Education*, *83*(5), 559–576.

Libarkin, J. (2008). Concept inventories in higher education science. *BOSE Conf*.

Libarkin, J. C., & Anderson, S. W. (2006). Development of the Geoscience Concept Inventory. *STEM Assessment Conference* (p. 148).

Lindell, R., & Olsen, J. P. (2002). Developing the lunar phases concept inventory. *Proceedings of the 2002 Physics Education Research Conference*.

Lindstrøm, C., & Sharma, M. (2010). Development of a Physics Goal Orientation Survey. *International Journal of Innovation in Science and Mathematics Education (formerly CAL-laborate International)*, *18*(2). Retrieved Jun6 6, 2012 from http://ojs-prod.library.usyd.edu.au/index.php/CAL/article/view/4063.

Martin, J. K., Mitchell, J., & Newell, T. (2004). Work in progress: Analysis of reliability of the Fluid Mechanics Concept Inventory. *Frontiers in Education, 2004. FIE 2004. 34th Annual* (p. F1F–3).

Meadows, G., & Wiesenmayer, R. L. (1999). Identifying and Addressing Students' Alternative Conceptions of the Causes of Global Warming: The Need for Cognitive Conflict. *Journal of Science Education and Technology*, *8*(3), 235–239.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *2*(3), 255–273. doi:10.1080/10705519509540013

Nottis, K. E., Prince, M. J., & Vigeant, M. A. (2010). Building an understanding of heat transfer concepts in undergraduate chemical engineering courses. *美中教育评论*, 1–8.

Nunnally, J. C. (1967). *Psychometric theory*. Tata McGraw-Hill Education.

Österlind, K. (2005). Concept formation in environmental education: 14-year olds' work on the intensified greenhouse effect and the depletion of the ozone layer. *International Journal of Science Education*, *27*(8), 891–908.

Pavelich, M., Jenkins, B., Birk, J., Bauer, R., & Krause, S. (2004). Development of a chemistry concept inventory for use in chemistry, materials and other engineering courses. *American Society of Engineering Education, Annual Conference* (pp. 3445–3452).

Plunkett, S., & Skamp, K. (1994). The ozone layer and hole: children's misconceptions. *Paper presented at the Australian Science Education Research Association*.

Pruneau, D., Liboiron, L., Vrain, E., Gravel, H., Bourque, W., & Langis, J. (2001). People's Ideas about Climate Change: A Source of Inspiration for the Creation of Educational Programs. *Canadian Journal of Environmental Education*, *6*, 121–138.

Rhoads, T. R., & Roedel, R. J. (1999). The wave concept inventory-a cognitive instrument based on Bloom's taxonomy. *Proceedings, 1999 Frontiers in Education Conference* (pp. 10–13).

Richardson, J. (2004). Concept inventories: Tools for uncovering STEM students' misconceptions. *Assessment and Education Research*, 19–26.

Richardson, J., Steif, P., Morgan, J., & Dantzler, J. (2003). Development of a concept inventory for strength of materials. *Frontiers in Education, Annual* (Vol. 1, pp. T3D29–33). Los Alamitos, CA, USA: IEEE Computer Society. doi:http://doi.ieeecomputersociety.org/10.1109/FIE.2003.1263342

Rowe, G., & Smaill, C. (2007). Development of an Electromagnetics Course-Concept Inventory-a work-in-progress.

Rye, J. A., Rubba, P. A., & Wiesenmayer, R. L. (1997). An investigation of middle school students' alternative conceptions of global warming. *International Journal of Science Education*, *19*(5), 527–551.

Schultz, L. (2009). *Understanding the Greenhouse Effect Using a Computer Model*. The University of Maine.

Shepardson, D. P., Niyogi, D., Choi, S., & Charusombat, U. (2009). Seventh grade students' conceptions of global warming and climate change. *Environmental Education Research*, *15*(5), 549–570.

Smith, M. K., Wood, W. B., & Knight, J. K. (2008). The Genetics Concept Assessment: A New Concept Inventory for Gauging Student Understanding of Genetics. *CBE-Life Sciences Education*, *7*(4), 422 –430. doi:10.1187/cbe.08-08-0045

Steif, P. S., & Dantzler, J. A. (2005). A statics concept inventory: development and psychometric analysis. *JOURNAL OF ENGINEERING EDUCATION-WASHINGTON-*, *94*(4), 363.

Stephanou, A. (2007). *The measurement of conceptual understanding in physics.* University of Melbourne.

Stone, A., Allen, K., Rhoads, T. R., Murphy, T. J., Shehab, R. L., & Saha, C. (2004). The statistics concept inventory: A pilot study. *Frontiers in Education, 2003. FIE 2003. 33rd Annual* (Vol. 1, p. T3D).

Stone, A. (2006). *A psychometric analysis of the statistics concept inventory*. University of Oklahoma.

Streveler, R. A., Olds, B. M., Miller, R. L., & Nelson, M. A. (2003). Using a Delphi study to identify the most difficult concepts for students to master in thermal and transport science. *Proceedings of the Annual Conference of the American Society for Engineering Education*.

Tongchai, A., Sharma, M. D., Johnston, I. D., Arayathanitkul, K., & Soankwan, C. (2009). Developing, Evaluating and Demonstrating the Use of a Conceptual Survey in Mechanical Waves. *International Journal of Science Education*, *31*(18), 2437–2457. doi:10.1080/09500690802389605

Wuttiprom, S., Sharma, M. D., Johnston, I. D., Chitaree, R., & Soankwan, C. (2009). Development and Use of a Conceptual Survey in Introductory Quantum Physics. *International Journal of Science Education*, *31*(5), 631–654. doi:10.1080/09500690701747226

Yeo, S., & Zadnik, M. (2001). Introductory thermal concept evaluation. *The Physics Teacher*, *39*, 496–504.