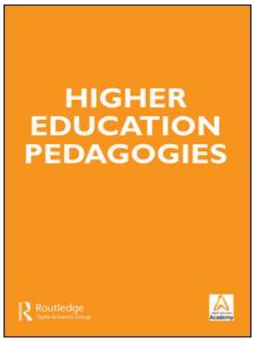
Using Concept Inventories to Measure Understanding

Journal Item



HIGHER EDUCATION PEDAGOGIES

Higher Education Pedagogies

ISSN: (Print) 2375-2696 (Online) Journal homepage: http://www.tandfonline.com/loi/rhep20

Using concept inventories to measure understanding

David Sands, Mark Parker, Holly Hedgeland, Sally Jordan & Ross Galloway

To cite this article: David Sands, Mark Parker, Holly Hedgeland, Sally Jordan & Ross Galloway (2018) Using concept inventories to measure understanding, Higher Education Pedagogies, 3:1, 60-69, DOI: 10.1080/23752696.2018.1433546

To link to this article: https://doi.org/10.1080/23752696.2018.1433546

Published online: 07 Feb 2018.

HIGHER EDUCATION PEDAGOGIES, 2018
VOL. 3, NO. 1, 60–69
https://doi.org/10.1080/23752696.2018.1433546

OPEN ACCESS

Using concept inventories to measure understanding

David Sands[a], Mark Parker[b], Holly Hedgeland[b], Sally Jordan[b] and Ross Galloway[c]

[a]Physics, School of Mathematics and Physical Sciences, University of Hull, Hull, UK; [b]School of Physical Sciences, The Open University, Milton Keynes, UK; [c]School of Physics and Astronomy, University of Edinburgh, Edinburgh, UK

ABSTRACT

Measuring understanding is notoriously difficult. Indeed, in formulating learning outcomes the word 'understanding' is usually avoided, but in the sciences, developing understanding is one of the main aims of instruction. Scientific knowledge is factual, having been tested against empirical observation and experimentation, but knowledge of facts alone is not enough. There are also models and theories containing complex ideas and interrelationships that must be understood, and considerable attention has been devoted across a range of scientific disciplines to measuring understanding. This case study will focus on one of the main tools employed: the concept inventory and in particular the Force Concept Inventory (FCI). The success of concept inventories in physics has spawned concept inventories in chemistry, biology, astronomy, materials science and maths, to name a few. We focus here on the FCI and ask how useful concept inventories are for evaluating learning gains. Finally, we report on recent work by the authors to extend conceptual testing beyond the multiple-choice format.

ARTICLE HISTORY
Received 1 September 2017
Revised 18 January 2018
Accepted 20 January 2018

KEYWORDS
Concept inventory; understanding; standardised tests

Introduction

There is no universally accepted definition of learning gain. HEFCE have defined it in broad terms as 'an attempt to measure the improvement in knowledge, skills, work-readiness and personal development made by students during their time spent in higher education' (England, 2017). How these things are measured is, of course, another matter and the reason for this special issue, but it should be apparent that this definition poses problems. By focusing on the whole time spent in higher education the implication is that learning gains defined in this way are best assessed at the end of a degree programme, but such is the range of topics that it would be difficult to identify some final, overarching assessment. The alternative is to measure these attributes by different means at discrete points along the journey from entry to graduation, but that's to assume that whatever assessment is made is durable and that the learning so measured is retained up to and beyond graduation. This can only be the case if the assessment accurately measures the state of knowledge: that is,

CONTACT David Sands ✉ d.sands@hull.ac.uk

it measures what students understand and can do without giving them chance to learn for the assessment.

The definition of learning gain adopted by McGrath et al. (McGrath, Guerin, Harte, Frearson, & Manville, 2015) is much simpler and more promising. Defined simply as, '"distance travelled" or learning acquired by students at two points of their academic career', this allows for assessment at identifiable points in the academic calendar. For an assessment of learning gain to be useful, though, there needs to be comparability between students and institutions and in describing the requirements that an assessment of learning gain must satisfy, McGrath et al. have in fact laid out the strategy employed by the Physics Education Research (PER) community many years ago when it set out to assess understanding. There is an argument, therefore, that the instruments developed out of Discipline-Based Educational Research in the sciences, and possibly further afield, for measuring students' grasp of concepts, so-called 'concept inventories', can be used to measure some of the discipline-specific aspects of learning gain. The primary purpose of this paper is to show, using the Force Concept Inventory (FCI) as an example, how these instruments lend themselves to measurements of learning gain. We report on recent work by the authors to overcome some of the disadvantages inherent in the multiple choice format by taking advantage of recent advances in technology to automate assessment of free-text answers. As physicists actively trying to assess understanding within our own discipline, our interest is in actively developing an instrument that captures the state of a student's knowledge. The context is clearly limited to the domain of the test, but in capturing the state of knowledge at given points we argue that concept inventories are suitable instruments for measuring learning gain.

Concept inventories: what are they and what do they measure?

A concept inventory is a multiple-choice research-level instrument designed to test students' conceptual understanding (Lindell, Peak, & Foster, 2007). Based on a number of key concepts from the subject (Jorion et al., 2015), each question, or item, has one correct answer and a number of incorrect answers, known as distractors, based on common student misconceptions (Sadler et al., 2009). The notion of a misconception, sometimes called an alternative conception (Caleon & Subramaniam, 2010) is crucial to measurements of conceptual understanding. Students develop surprisingly similar notions and apply them, sometimes consistently and sometimes inconsistently, to answer questions going beyond declarative knowledge and requiring qualitative reasoning. Identifying misconceptions is therefore central to characterising a student's state of understanding and is usually one of the intended outcomes of a concept inventory. The key to constructing an effective concept inventory lies in the selection of good questions with appropriate distractors (Dick-Perez, Luxford, Windus, & Holme, 2016).

The most common use of concept inventories is to test the effectiveness of a particular pedagogical practice in altering those alternative conceptions (Porter, Taylor, & Webb, 2014). Students are tested on the concept inventory before instruction (known in the literature as the 'pre-test') and again after instruction (known in the literature as the 'post-test'). The pre-test and post-test scores from across the student body are then compared in order to gauge the effectiveness of the teaching methods (Bailey, Johnson, Prather, & Slater, 2012). The results of this comparison can then be used to guide future instruction by adapting the teaching to address the conceptual deficiencies revealed by the tests.

Pedagogical development and learning gain are closely related: if pedagogy is ineffective there should be little or no learning gain. That is not to say that students will not have learnt something, but it might not be what is intended. Much depends on what is being assessed. By definition, a student who passes an exam will have learnt something, but PER has shown that being able to pass an exam does not necessarily equate to developing an understanding. Concept inventories are intended to assess understanding and the implication of the above is that they assess different things from conventional exams. According to Epstein (2013), 'There does not seem to be a universally accepted definition of what constitutes a concept inventory' (p. 1019), but he described them as, 'tests of the *most basic* conceptual comprehension of [the] foundations of a subject and not of computation[al] skill. They are quite different from final exams and make no pretence of testing everything in a course' (p. 1019). It is this last point that is really important: concept inventories are different from final exams or unseen summative assessments for four main reasons:

(1) Summative assessments are taken once, so there is no baseline against which the performance can be compared.
(2) Students usually have advance notice of a summative assessment and therefore have time to prepare, leading to the possibility that much of the learning demonstrated in the assessment is not real or lasting but simply sufficient for the assessment.
(3) Unlike concept inventories, summative assessments, particularly time-limited unseen examinations, often test mathematical facility or declarative knowledge.
(4) Summative assessments, especially final exams, are high-stakes.

Concept inventories on the other hand are low-stakes tests usually taken twice within the instruction sequence. As such, concept inventories fulfil a number of requirements for assessing learning gain. As described by McGrath et al. (2015) assessments of learning gain must be:

(1) *Valid*. Tests are validated through research to ensure they measure what they claim to measure.
(2) *Standardised*. In so far as the same test is used on different students in different institutions, the tests allow for meaningful comparisons of different students' understanding of the concept in question.
(3) *Longitudinal*. The same test is used at two different points in time and allows for a meaningful assessment of gain by the class as a whole as well as individual students, if the different responses can be matched.

Concept inventories fulfil all these requirements, but it is their ability to measure understanding rather than declarative knowledge that makes them potentially effective measures of learning gain.

The FCI was the first concept inventory to be published and in describing the development of the Calculus Concept Inventory, Epstein paid due attention to the fore-runners from physics:

> All of them [concept inventories] trace their roots to the MDT [Mechanics Diagnostic Test] and FCI [Force Concept Inventory] in physics, and there is general agreement that physics education is ahead of other disciplines in the use of concept tests as measures of teaching effectiveness and in the development of programs that show much improved gain. (p. 1019)

What prompted the development of such instruments in physics was the observation that post-physics course students in MIT demonstrated 'unnerving similarity … in non-Newtonian expectations' to five-year olds (di Sessa, 1987). This is a remarkable observation and highlights the difficulty in measuring exactly what students have learned over the course of their degree. Students who had taken a course specifically designed to teach them Newtonian mechanics, and presumably passed the assessments, not only had no real grasp of the foundations but in fact possessed incorrect ideas developed in childhood that persisted through formal education. It is a finding echoed throughout the history of PER and summarised by McDermott (2001) in her Oersted medal winning lecture of 2001: 'Certain conceptual difficulties are not overcome by traditional instruction. (Advanced study may not increase understanding of basic concepts.)' (p. 1131). Moreover, a 'coherent conceptual framework is not typically an outcome of traditional instruction' (p. 1133) and 'growth in reasoning ability often does not result from traditional instruction' (p. 1132). Yet, like the students at MIT mentioned by di Sessa, these students will have faced some final assessments. Whatever these assessments were testing, it was not these attributes.

It is against this background that the FCI, the first, and most famous, concept inventory, was developed (Hestenes, Wells, & Swackhamer, 1992). It was not without its critics, however. Huffman and Heller (1995) in particular argued that the FCI was a test of mastery of various contexts and content relating to force, not a test of the force concept itself. More recently, Wang and Bao (2010) analysed the FCI using item response theory and reported an uneven performance of some items or questions. Hestenes and Halloun (1995) maintained in response to Huffman and Heller that the FCI does indeed measure conceptual understanding, and Hake's now famous paper (Hake, 1998) showing how data from the FCI revealed that interactive engagement (IE) within teaching was effective for teaching Newtonian mechanics does indeed suggest that it measures something useful.

As a result of Hake's findings, which led to large-scale reform in physics education (Reed-Rhoads & Imbrie, 2008), the FCI itself is partially credited with much of the reform that has taken place in physics education over the past two decades (Scott, Peter, & Harlow, 2012). The idea of a test of qualitative reasoning having something useful to say about the state of students' learning is now well established. The FCI itself has been used widely throughout the world (Lasry, Rosenfield, Dedic, Dahan, & Reshef, 2011) and though there are disagreements about what precisely it measure (Wallace & Bailey, 2010) it is generally agreed in the community that the FCI does indeed measure something useful. If the FCI is indeed flawed, this only points to the difficulty in designing effective tests of understanding and hence learning gain.

The FCI's success and subsequent educational reforms in physics led to the development of concept inventories in other areas of physics and astrophysics as well as a wide range of STEM subjects such as chemistry, biology, mathematics, geoscience, engineering and more recently computer science. Rather than provide a list of references, which will inevitably be incomplete and by no means systematic, we leave it to the interested reader to identify to identify instruments within their own domain. In addition, concept inventories have been developed extensively for use in secondary schools, as illustrated by the work of Caleon and Subramaniam (2010) and references therein.

It is generally acknowledged, at least in higher education, that all concept inventories can trace their roots to the FCI, and each follows a similar, iterative pattern of development (Porter et al., 2014): gather together concepts that are to be tested using the inventory, come up with the questions and responses, either by consulting the literature, using student essay and interview responses (the preferred approach of Bailey et al., 2012), or by using the judgement of experts, and then roll out the pilot test and gather data. This sequence is iterated until the concept inventory attains the desired level of quality.

It would be misleading to suggest that all concept inventories have the same level of validity and reliability. We have already alluded to concerns over exactly what the FCI measures and it is possible to find similar concerns raised against other concept inventories in other fields. This is obviously important if concept inventories in general, and not just the FCI, are to be used to measure learning gain. Again, rather than try to be exhaustive we draw on the work of the physics community to illustrate the difficulty in developing valid instruments. Lindell et al. (2007) compared the development steps of twelve different inventories used at North Carolina State University. The concept domains tested by these inventories were identified by three different methods: from student-oriented investigations, suggestions from the literature and expert knowledge of the discipline. In consequence, in three of the twelve, the distractors were based on student understanding alone and in three others, they were based on expert understanding alone. In two of the instruments, the distractors were based on a combination of these two methods and four didn't specify how the distractors were identified. Crucially, only two of the twelve were found to have distractors that could correctly identify student misconceptions, suggesting perhaps that not all concept inventories will effectively characterise students' understanding.

In looking at whether the items differentiated between high-performing and low-performing students and the proportion of students who could correctly answer a question, Lindell et al. (2007) also found significant deficiencies. These authors concluded that none of the inventories had been validated sufficiently rigorously, but in nine of the twelve there was sufficient information about them to conclude that they were reliable. The authors suggested the need for a formal classification for concept inventories and an accepted standardised procedure for their development. Such a classification system would probably be necessary if these instruments are to be used in measuring learning gain. This variability is further emphasised by the analysis of Jorion et al. (2015), who conducted a review of three concept inventories using a range of approaches. Of the three analysed, two could be used to test students' overall understanding, one could test for understanding of specific concepts, but none of them could explain students' choice of misconceptions and errors.

Concept inventories and learning gain

Taking the notion of learning gain as 'distance travelled', concept inventories would appear to be well placed to act as effective instruments. In a sense, this is what they already do, but perhaps with varying levels of effectiveness. The FCI has been especially effective in showing that some pedagogies lead to better learning than others. A direct example illustrates the point. One of us (Sands, 2010) taught Newtonian mechanics through the medium of the 3-D visual programming environment, VPython. Based on the work of Ruth Chabay at North Carolina (see e.g. Chabay & Sherwood, 2008), the idea was for students to create objects on screen, which can be done very easily with VPython, assign properties to them and move them according to the laws of Newtonian mechanics. The visual motion allows for direct observation of the evolution of the motion over time, thereby giving instant feedback on both the model and the Newtonian concepts. However, the FCI revealed no discernible change in the conceptual understanding and further investigation revealed that many students focussed on the production of a computer programme rather than a physics model. By dispensing with programming and concentrating on the process of building mathematical models themselves, significant and consistent improvement in conceptual understanding was recorded (Sands & Marchant, 2012). One pedagogy was effective in delivering learning gain whilst another was not.

The effectiveness of the FCI as a measurement of learning gain derives from its intention to measure student thinking rather than declarative knowledge about laws and principles and procedural knowledge about mathematics. There is also an emphasis on misconceptions, but the FCI might well be unique in this respect. The term, 'misconception' doesn't just mean a misunderstanding, it is taken to imply an alternative view of the world. Everybody holds, or has held, such misconceptions about mechanics at some time in their life because everybody has had direct experience of force and motion from an early age. Non-Newtonian causal explanations developed to account for those experiences carry through into adulthood and into higher education, with research revealing them to be remarkably common among different individuals. The FCI can meaningfully test for these alternative views, but in other subjects, and even in other topics within physics, there might not necessarily be any reason to expect such characteristic alternative conceptions, as opposed to simply misunderstandings of taught material. The finding by Lindell et al. that not all concept inventories effectively test for misconceptions is probably not that important because there might be no inherent reason why they should. What is important is whether the student has moved on from the initial position and how we can measure this. In short, how well the test functions as a measurement of understanding will determine to a large extent how effective it is a test of learning gain.

Whether concept inventories could be used across the content of a degree programme is open to question. As with the FCI, many instruments have been designed for use at relatively junior undergraduate levels in foundational topics where it might be expected that there is sufficient common content across similar degree programmes in different universities to enable such instruments to function as standardised tests of learning gain. However, not only is it rare to find concept inventories for more advanced and more specialised topics, but the diversity of content might prohibit their development. This would necessitate a different approach at these levels. Notwithstanding this limitation, it would seem safe to conclude that concept inventories have the potential to play a useful role in measuring learning gain, but whether the format, and hence the functioning as a test of learning, can be improved upon is as yet an unanswered question.

Concept inventories are usually multiple choice questionnaires (MCQ) for ease of deployment to large numbers of students across institutions. The MCQ format has its limitations. Caleon and Subramaniam (2010) developed a three-tier test for use at secondary level based on two-tier tests, which seem to be popular at secondary level since Treagust's work in the late 1980s (Haslam & Treagust, 1987). The two-tiers test, respectively, propositional statements and the scientific reasoning behind the propositions, but, as Caleon and Subramaniam point out, '… a two-tier test … cannot differentiate mistakes due to lack of knowledge from mistakes due to existence of ACs [alternative conceptions]; conversely, it cannot differentiate correct responses due to adequate understanding from those due to guessing …' (p 941). The third tier tests for confidence in the answer and is intended to distinguish between guessing and intentional choices.

Caleon and Subramaniam make no distinction between concept testing at university and at secondary level, but their test, and indeed other two-tier tests used at secondary level, operate in a different way from the FCI, for example. Some questions in the FCI might be regarded as testing content knowledge, but in others students are explicitly required to reason about a problem. In a two-tier test, however, students have to *select* the scientific reason underlying the propositional statement tested in the first tier. Without undertaking a comprehensive survey, we cannot say how general is this difference, but it is plausible to suggest that it reflects the difference between what secondary-level students and university-level students might be expected to do. In other words, a focus on content knowledge might be appropriate at secondary level, but less so at the tertiary level where the focus is on conceptual understanding and the application of knowledge. This might suggest that two and three-tier tests would not represent a useful development for measuring learning gain. Moreover, each tier essentially adds a question to the test, which inevitably reduces the total number of questions that can be asked.

The MCQ format suffers from another, perhaps less obvious difficulty: it is not always the case that students choosing a right or wrong answer do so for the reasons we might think. Sands (2013) has shown that for certain questions in the FCI that students might well have all the necessary knowledge to answer a question correctly, but still choose an incorrect option, thereby appearing to act irrationally. Students' thinking can be quite complex and context has been suggested to play a role in students' poor performance in the FCI (Stewart, Griffin, & Stewart, 2007); (Wang & Bao, 2010). Ideally, a test of learning gain would eliminate such difficulties, as would an ideal concept inventory. Our concern is that such difficulties might arise from the very fact of having to choose from a prescribed set of answers.

We have identified within our research a clear need to go beyond the MCQ format because we cannot be sure from a wrong choice on a MCQ the extent of any misunderstandings. A good example is provided by question 15 on the FCI, which tests understanding of Newton's 3rd law. The question is posed in terms of a car pushing a truck whilst both are accelerating and asks which is greater: the force exerted by the car on the truck or the truck on the car? Newton's 3rd law requires them to be equal, but the great majority of students with no prior university instruction and only A-level knowledge choose the option in which the force from the car on the truck is greatest. Further research revealed that the majority of students know Newton's 3rd law in as much as they can state it correctly even months after their last A-level contact, but they choose to apply Newton's 2nd law because the system is accelerating and there must be a net force on the car. When shown through the process of building a mathematical model (Sands & Marchant, 2012) how Newton's 3rd law applies in this situation, the majority of students get this question correct on post-instruction testing. It is not that they have learnt Newton's 3rd law, which would be an obvious conclusion on the basis of the FCI alone, or overcome an alternative conception, but they have understood how the 3rd law applies even when a net force must exist in one direction. This is a measurable learning gain, but it is not revealed by the FCI in isolation.

We are researching a format in which students construct their own answers. This overcomes the difficulty that students are constrained in their choice of answers, and we can be confident that any conceptual difficulties genuinely reflect the state of a student's understanding. We have translated the FCI as far as we can into a version that requires free-text responses, though we are exploring other formats for questions involving trajectories or other answers that are not easy to describe in words. For the question on Newton's 3rd law just discussed, two very interesting examples of complex thinking that is not captured within the multiple-choice options are: ‘Forces are the same, but the truck will accelerate at a slower pace due to more mass (m1a1 = m2a2, N’s 3rd Law)’ and ‘The force exerted by the truck gradually decreases as the system accelerates’. Even though the magnitudes of the forces have been correctly identified as being equal in the first answer, and the student has essentially answered the question correctly, the notion that two bodies in contact can accelerate differently and remain in contact is clearly wrong. Not only is there a failure to reason correctly about the forces acting as the force driving the car is neglected from the equation, but there is also a significant failure to grasp the nature of acceleration. Encapsulating this alternative in a question without perplexing students for whom the notion is nonsensical would be a significant challenge, but in a free-text response this misunderstanding is revealed quite naturally. The reasoning behind the second answer is not at all clear at this stage, but it represents an interesting view not hitherto recognised within the canon of non-Newtonian misconceptions.

Another odd view revealed by free-text answers that would not show up in a MCQ test concerns the force of gravity and its manifestation as weight. Allowing for a certain looseness of language, these are the same thing, but a number of students explicitly identified them as separate. The question asked: ‘A stone is dropped from the roof of a single story building to the surface of the earth. State what forces are acting on the stone when it is in flight’. Answers included, ‘Friction, gravity, weight’, ‘gravity and weight are both acting on top of stone’, ‘Gravitational field strength, weight and frictional force’ and ‘Gravity, Air resistance and the weight force’. Again, it would be difficult to encapsulate this in a MCQ test in a way that a student who recognised weight and gravity as essentially the same would not think it a trick question. These two examples illustrate that with free-test answers, we are able to assess an initial position and whether a student has moved on. We suggest that this represents a real measurement of learning gain.

The move to free-text responses raises a whole host of research questions which are as yet unanswered. Some of them revolve around the implications of the answers, as illustrated above, and others around the format and the technology. For example, automatic assessment becomes even more important in these circumstances in order to ensure not only consistency of marking, especially if the test is to become standardised, but also ease of deployment to large numbers of students. Manually assessing around 30 free-text answers per student is by no means a simple task so we are using pattern-matching question types developed at the Open University (Jordan, 2012). These work best if the answers are short and we don’t yet know whether the cognitive processes involved in thinking about a short sentence are different from those involved in choosing from a set of fixed answers. If they are, then the implications for conceptual testing and measuring learning gain are profound: measured gains might depend not only on context but also on the type of question asked.

In conclusion, we have demonstrated that concept inventories are a promising tool for measuring learning gains in specific areas of the curriculum. Designed primarily to test the effectiveness of a particular pedagogy, the tests necessarily measure the kind of developments in students that a test of learning gain would also measure. However, concept inventories are not perfect tools, despite the huge amount of effort devoted to their development, and we suggest that true measurement of conceptual understanding, and by implication learning gain, is a complicated and as yet ill-defined process. We have described our own approach to developing the format of the concept inventory and suggest that free-text responses are potentially effective for assessing learning gains in a way that the conventional MCQ does not.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

David Sands http://orcid.org/0000-0003-1083-604X
Sally Jordan http://orcid.org/0000-0003-0770-1443

References

Bailey, J.M., Johnson, B., Prather, E.E., & Slater, T.F. (2012). Development and validation of the star properties concept inventory. *International Journal of Science Education, 34*(14), 2257–2286.

Caleon, I., & Subramaniam, R. (2010). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education, 32*(7), 939–961.

Chabay, R., & Sherwood, B. (2008). Computational physics in the introductory calculus-based course. *American Journal of Physics* 76(4), 307–313. doi:10.1119/1.2835054

di Sessa, A. (1987). The third revolution in computers and education. *Journal of Research in Science Teaching, 24*(4), 343–367. doi:10.1002/tea.3660240407

Dick-Perez, M., Luxford, C.J., Windus, T.L., & Holme, T. (2016). A quantum chemistry concept inventory for physical chemistry classes. *Journal of Chemical Education, 93*(4), 605–612.

England, H. (2017). *Learning gain – Higher Education Funding Council for England*. Hefce.ac.uk. Retrieved April 15, 2017, from http://www.hefce.ac.uk/lt/lg/

Epstein, J. (2013). The calculus concept inventory – Measurement of the effect of teaching methodology in mathematics. *Notices of the American Mathematical Society, 60*(8), 1018–1026.

Hake, R.R. (1998). Interactive-engagement vs traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*, 64–74.

Haslam, F., & Treagust, D.F. (1987). Diagnosing secondary students' misconceptions of photosynthesis and respiration in plants using a two-tier multiple choice instrument. *Journal of Biological Education, 21*(3), 203–211.

Hestenes, D., & Halloun, I. (1995). Interpreting the Force Concept Inventory: A response to March 1995 critique by Huffman and Heller. *The Physics Teacher, 33*, 502–506.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher, 30*, 141–158.

Huffman, D., & Heller, P. (1995). What does the Force Concept Inventory actually measure? *The Physics Teacher, 33*, 138–143.

Jordan, S. (2012). Student engagement with assessment and feedback: Some lessons from short-answer free-text e-assessment questions. *Computers & Education, 58*(2), 818–834.

Jorion, N., Gane, B.D., James, K., Schroeder, L., Dibello, L.V., & Pellerino, J.W. (2015). An analytic framework for evaluating the validity of concept inventory claims. *Journal of Engineering Education, 104*(4), 454–496.

Lasry, N., Rosenfield, S., Dedic, H., Dahan, A., & Reshef, O. (2011). The puzzling reliability of the Force Concept Inventory. *American Journal of Physics, 79*(9), 909–912.

Lindell, R.S., Peak, E., & Foster, T.M. (2007). *Are they all created equal? A comparison of different concept inventory development methodologies*.AIP conference proceedings 883, (pp.14–17). Syracuse, New York. https://doi.org/10.1063/1.2508680

Lillian Christie McDermott. 2001. Oersted Medal Lecture 2001: "Physics Education Research—The Key to Student Learning". American Journal of Physics. 69(11) 1127-1137.

McGrath, C., Guerin, B., Harte, E., Frearson, M., & Manville, C. (2015). *Learning gain in higher education* (1st ed., p. 7). Santa Monica, CA: RAND Corporation. Retrieved from https://www.rand.org/pubs/research_reports/RR996.html

Porter, L., Taylor ,C., & Webb, K. (June 23 - 25, 2014). *Leveraging open source principles for flexible concept inventory development*. Proceedings of the 2014 conference on innovation & technology in computer science education, (pp. 243–248). Uppsala, Sweden.

Reed-Rhoads, T., & Imbrie, P.K. (2008). Concept inventories in engineering education. Paper presented at the National Research Council's Workshop on Linking Evidence to Promising Practices in STEM Undergraduate Education. (pp. 13–14). Washington, DC. October. Accessed from: https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_072625.pdf

Sadler, P.M., Coyle, H., Miller, J.L., Cook-Smith, N., Dussault, M., & Gould, R.R. (2009). The astronomy and space science concept inventory: Development and validation of assessment instruments aligned with the K-12 national science standards. *Astronomy Education Review, 8* (1). doi:10.3847/AER2009024.

Sands, D. (2013, August). *Evidence for the applicability of dual processing theory in physics problem solving*. Proceedings ICPE-EPEC, Prague. Retrieved from http://www.icpe2013.org/uploads/ICPE-EPEC_2013_ConferenceProceedings.pdf

Sands, D. (2010). First year mechanics taught through modelling in VPython. *New Directions, 6*, 47–50 Retrieved from https://journals.le.ac.uk/ojs1/index.php/new-directions/article/view/383/388

Sands, D., & Marchant, A.L. (2012). Enhanced conceptual understanding in first year mechanics through modelling. *New Directions, 8*, 22–26 Retrieved from https://journals.le.ac.uk/ojs1/index.php/new-directions/article/view/490/487

Scott, J., Peter, M., & Harlow, A.(20-23 August, 2012). *An electronics threshold-concept inventory*. Proceedings of IEEE international conference on teaching, assessment, and learning for engineering. Hong Kong. doi:10.1109/TALE.2012.6360366

Stewart, J., Griffin, H., & Stewart, G. (2007). Context sensitivity in the Force Concept Inventory. *Physical Review Special Topics – Physics Education Research, 3*(1). https://journals.aps.org/prper/abstract/10.1103/PhysRevSTPER.3.010102.

Wallace, C.S., & Bailey ,J.M. (2010). Do concept inventories actually measure anything?*Astronomy Education Review, 9* (1). http://access.portico.org/Portico/#!journalAUSimpleView/tab=PDF?cs=ISSN_15391515?ct=E-Journal%20Content?auId=ark:/27927/pgg3ztfdrhv

Wang, J., & Bao, L. (2010). Analyzing Force Concept Inventory with item response theory. *American Journal of Physics, 78*(10), 1064–1070.