Diagnostic tests for the physical sciences: A brief review

Simon Bates* and Ross Galloway
Physics Education Research
School of Physics and Astronomy
University of Edinburgh
Edinburgh
EH9 3JZ
***s.p.bates@ed.ac.uk**

Abstract
We present a review of diagnostic testing in the physical sciences. We cover the motivation for using such instruments and their historical development via a case study of probably the most cited and influential test instrument and application: the Force Concept Inventory, developed in the early 1990s by Hestenes and co-workers, and its use to quantify learning gains from different instructional methodologies by Richard Hake. We then present an overview of the process of creation and validation of such instruments, and highlight the results from studies that have made use of some of the many instruments available in the literature. We conclude with a short summary of our own recent work to develop a diagnostic test of data handling skills of physical science undergraduates.

The aim of this review is to present a brief overview of some of the tests that exist within the literature and have been developed and deployed to test attributes from broad 'scientific thinking' ability to conceptual understanding of specific areas of physical science.

1. Introduction and background

The last twenty years or so have seen considerable effort directed towards the development, validation and application of diagnostic tests in the physical sciences: standardised testing instruments designed to yield a robust, reliable and quantitative measure of student understanding on a particular topic or subject area. They usually take the form of multiple-choice questions (MCQs)[1,2] designed to test conceptual understanding as opposed to bald factual recall. Many of the tests originate from the Physics Education Research effort in the US, but the principles (and in some cases, the concepts they examine) are relevant more broadly across other science disciplines, especially chemistry.

A standardised, expertly-validated diagnostic instrument that is capable of yielding deep insights into the conceptual understanding (or otherwise!) of students at a particular stage in their studies holds an obvious appeal. The motivation for their use is succinctly embodied in a quote from the late David Ausubel:

> *"If I had to reduce all of educational psychology to just one principle, I would say this: The most important single factor influencing learning is what the learner already knows. Ascertain this and teach him accordingly."* [3]

This is one of the main use scenarios for such instruments: the assessment of knowledge and understanding, often prior to commencing further study. This is the ethos behind the Open University 'Are you ready for….?' student self-assessments as course precursors[4]. Of equal validity is to look at 'residual' understanding long after explicit teaching, to differentiate real conceptual understanding, committed to long term memory as opposed to short term recall[5]. Another widespread use of such instruments is both pre- and post-instruction (often with the same test, possibly an isomorphic one that tackles the same concepts with different questions). The most widely-cited example of this is Richard Hake's 1998 study[6] of more than 6000 students' conceptual understanding of classical mechanics, using the Force Concept Inventory (FCI) devised by Halloun and Hestenes[7], which is covered in more detail in the following section.

The aim of this review is to present a brief overview of some of the tests that exist within the literature and have been developed and deployed to test attributes from broad 'scientific thinking' ability to conceptual understanding of specific areas of physical science. In addition, we will highlight certain areas of application of these instruments and the findings that they have yielded. We cannot be completely comprehensive in the space available, so a 'broad brush' approach is necessarily adopted. We hope the review will be of value to those colleagues dipping their toes into this arena for the first time, as well as more experienced staff who want a more detailed account of aspects of instrument creation and validation. The paper is organised as follows: the next section presents a case study of one particular test and its most-cited application: Hake's study

of conceptual understanding of mechanics using the FCI[6]. This is a seminal study that set the standard to which many, if not all, subsequent investigations have aspired. We then change tack slightly and consider the process of devising, validating and testing an instrument. Once again, we make use of a particular instrument to exemplify the procedure: the Basic Electricity and Magnetism Assessment, by Beichner and co-workers. Devising and validating an instrument is a time-consuming process, and many instruments have already reached this level of maturity and can be used by staff 'off the shelf'. The third section presents an overview of some of the available instruments and their applications (with links to others). Finally, we conclude with details of some of our own work to devise and validate a diagnostic instrument to test data handling skills of physical science undergraduates.

2. A case study: the Force Concept Inventory

The development of the FCI can be traced back to the Mechanics Diagnostic Test, first published in 1985, based on the dissertation research of Ibrahim Halloun[7]. It comprises MCQs covering conceptual topics in Newtonian mechanics, a subject all Physics students entering University will have had considerable exposure to, and in which will have solved a large number of 'problems'[9]. To some staff, the test items look simple and they deliver it to students confident of high scores, yet are usually surprised by the results. The most well-cited example is Eric Mazur's experience at Harvard, where it was noted that students could solve complex quantitative problems in mechanics, yet fail to correctly answer some of the (supposedly easier) conceptual questions on the FCI. This experience led Mazur to develop the instructional methodology of Peer Instruction[10] (the book of the same name includes a slightly revised version of the test), now widely adopted as a tool for interactive engagement and enhanced conceptual understanding. This methodology, plus the widespread introduction of an effective mediating technology (in the form electronic voting system handsets in lectures), illustrates just how far the FCI ripples have spread.

The FCI describes six 'conceptual dimensions' (kinematics, Newton's three laws, kinds of forces and superposition of forces) from which a taxonomy of student misconceptions (or 'alternate conceptions') has been derived. A much more detailed analysis of these dimensions is presented in the original references and elsewhere[11]. The key research findings that followed from implementation of the test by the authors suggested significance for undergraduate teaching and learning that went far beyond the content topic of Newtonian mechanics. There appeared to be little correlation between FCI scores and mathematical ability or socioeconomic level, and scores obtained prior to teaching were uniformly low. There appeared to be virtually no correlation between FCI test scores after teaching and teacher competence.

Many of these findings were convincingly reconfirmed and extended by Richard Hake's study of over 6000 students' results from taking the FCI prior to and after courses in classical mechanics (often called a 'pre- / post-' testing methodology). Hake set out to try and understand and quantify the effects of different types of two broad categories of instruction on conceptual understanding. The first of these categories was the 'traditional' instruction methods, characterised by largely didactic lectures requiring little student involvement, recipe-based laboratories and algorithmic problems for assessment. The second methodology Hake termed 'interactive engagement' (IE), a broadly defined umbrella term which is characterised by engagement of "*students in heads-on (always) and hands-on (usually) activities which yield immediate feedback through discussion with peers and/or instructors*"[6].

The results are remarkable: Figure 1 shows a summary of all data collected, comprising a total of 6542 students (2048 enrolled in 14 courses characterised as 'traditional' delivery, the remainder in 48 IE-type courses across a range of types of educational institution in the US). The figure plots class average pre-test score on the abscissa ('<pretest>') against percentage gain from the post-test on the ordinate. Each data point represents a given class / cohort and, reassuringly, all

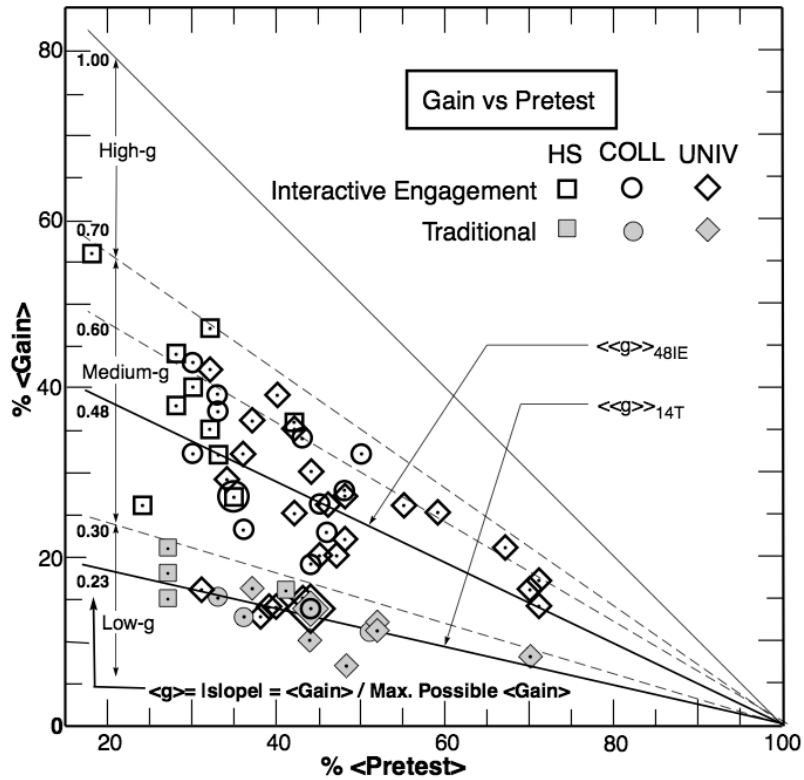

Figure 1: Post-test improvement as a function of pre-test score for a sample of 6542 students taking the Force Concept Inventory test of conceptual understanding in classical mechanics. Reproduced from reference 6.

classes show a positive percentage gain, indicating average performance improved post-instruction. A key quantity used to characterise the learning gains between pre- and post-instruction testing is the cohort-averaged normalised gain, $\langle g \rangle$[6,12,13], calculated as:

$$\langle g \rangle = \frac{\langle S_f \rangle - \langle S_i \rangle}{100 - \langle S_i \rangle}$$

Where $\langle S_f \rangle$ and $\langle S_i \rangle$ are the post and pre class averages as percentages, respectively. Hake further characterised courses on the basis of these average normalised gains, where 'high-g' courses have ($\langle g \rangle$) > 0.7 (though none of his data fell in this range); 'medium-g' where 0.7 > ($\langle g \rangle$) > 0.3 and 'low-g' where ($\langle g \rangle$) < 0.3. Most strikingly, all the traditional courses fell within the lowest of the three bands and most of the IE courses in the medium-g region, albeit with a broader spread. The mean values of the mean normalised gains for a particular type of instruction, $\langle\langle g \rangle\rangle_{48IE}$ and $\langle\langle g \rangle\rangle_{14T}$, as indicated on Figure 1, differed by a factor of 2. In other words, the IE courses were, on average, about twice as effective in enhancing conceptual understanding of the material as traditional courses. These gains, as Hake remarks in the original paper, offer strong evidence of one route to a solution to Bloom's '2 sigma' problem[14], the challenge to find instructional methodologies for group instruction that are as effective as individual tutoring. A closer inspection of the raw data is provided by Hake's companion papers from around the same time[15].

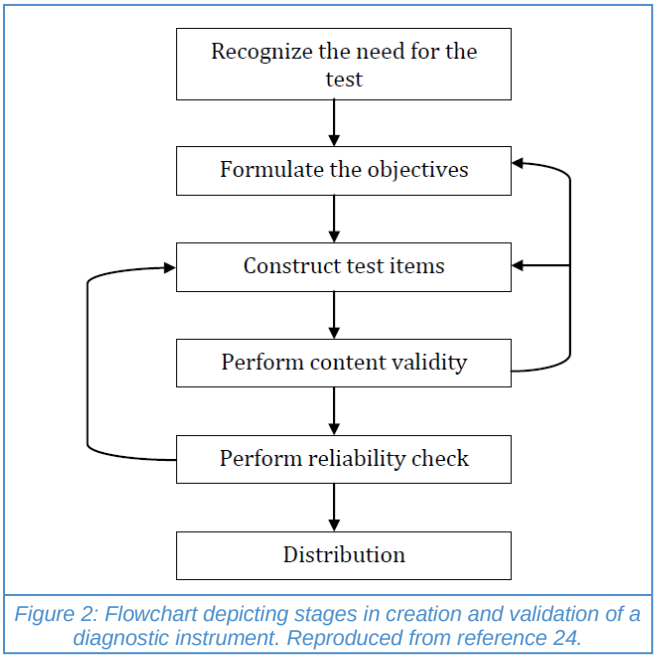

Figure 2: Flowchart depicting stages in creation and validation of a diagnostic instrument. Reproduced from reference 24.

Hake's study sparked extensive and widespread debate: critiques and responses to critiques abound in the literature. The interested reader is directed towards a few springboard papers[11,16]: there are many others. Scrutiny of these reveals subtleties and complexities: some IE courses achieve ($\langle g \rangle$) $\langle 0.3$; there is often a very large spread in *g* values for students on a given course; the fact that traditional courses in the survey produced low $\langle g \rangle$ values does not rule out the fact that some traditionally delivered courses may yield medium-*g* scores. Furthermore, the FCI is a particular type of assessment: Mahajan has observed that even students who score very well on the pre-test can have significant difficulties answering free response problems of a conceptual nature or that require estimation skills[17]. However, laying this debate to one side, it is abundantly clear that this paper has had an enormous and lasting effect. Its impact has been felt both within the physical sciences and across many other disciplines and its findings have been used as the basis for a great deal of curriculum change and reform. At the time of writing, it has been cited 343 times.

Our own experience in Physics and Astronomy at Edinburgh with the FCI stretches back a mere 6 years, influenced by Hake's guidelines for administering the test[18]. Since then, we have administered it consistently pre- and post- instruction. We have done this both as a measure of student conceptual understanding in the topic area at the entry and exit points of the course, but also as a measure of effectiveness of the teaching on the course as we have progressively incorporated IE elements more consistently. These have included studio based teaching approaches[19], electronic voting systems[20] and elements of Peer Instruction[10]. In terms of the students' understanding, our results show many similarities with previously published studies: on entry, average conceptual understanding of the cohort is significantly below the 60% level, identified by Hestenes[7] as the 'entry threshold'. Below this threshold, student understanding of the concepts is deemed to be insufficient for effective problem solving. In terms of measuring the effectiveness of the instruction, we have seen $\langle g \rangle$ rise from 0.3 to consistently around or above 0.5, with a substantial fraction (but by no means all) of the cohort attaining the 'mastery threshold' of 85% on post-instruction testing. What is striking is that the performance on many of the questions, in terms of not only percentage of students choosing the correct answers, but those who choose particular distracters, is almost completely invariant over time. Some questions have a large and consistent fraction of the cohort choosing the same wrong answer year after year: we call these 'banana-skin' questions, as successive year of students seem to slip up on them in the pre-test. The construction of the questions is such that one is able to directly ascertain the world-view that students are operating within, for example the pre-Newtonian conception that 'motion must imply a force'.

Before we conclude this section on the FCI, we should also point out some of its close relatives. The closest of these is the Mechanics Baseline Test (MBT)[21], which focuses more on processes rather than concepts. Both the FCI and MBT have been further developed by one of the original authors (Halloun) in work to devise basic inventories of concepts and processes (IBC-Mechanics and IBP-Mechanics, respectively)[21]. Other instruments include the Force and Motion Conceptual Evaluation, described along with the effects of research-based active learning strategies to improve conceptual understanding[23].

3. Creation and validation of diagnostic test instruments
Generating a diagnostic test which is sufficiently robust as to provide an effective measure of student performance is necessarily an involved task. Here we are able to give only a brief overview; for the interested reader, Engelhardt gives a very detailed and readable account of the entire test creation process from start to finish[24]. In essence, the process involves multiple stages, many of which may feature multiple iterations, and which can occupy many months to years. Beichner[25] has succinctly encapsulated the process in a flowchart, shown in Figure 2.

As this figure shows, creation of a diagnostic test instrument involves much more than just writing the test questions. The starting point should always be the identification of a specific area which requires diagnostic investigation: 'chemistry' is likely to be too broad, whereas 'spectroscopic notation in quantum mechanics' is probably too narrow. Frequently, diagnostic tests are structured to cover areas which would naturally fall within a single 'lecture course' unit in the higher education setting. Before embarking further on test creation, at this stage it is worth checking the literature to make sure a suitable instrument does not already exist: diagnostic testing in the physical sciences is a rapidly growing field. (A non-exhaustive summary of some existing instruments is given in section 4 of this review.)

Having identified the desired field for the instrument, the next requirement is to establish the learning objectives it is intended to assess. These should be framed in terms of student competency (i.e. "Students should be able to…"), and are often focussed on areas where instructors find that traditionally there are widespread difficulties or alternate conceptions. It is important not to be too ambitious here: Engelhardt reports recommendations of 5 to as many as 20 questions per objective, and suggests a minimum of 3 questions per objective for 'low-stakes' tests such as diagnostic instruments[24]. Thus, to prevent the test becoming unmanageably long, there are a practical maximum number of objectives, probably at most ten.

To mitigate the chances of wasted effort at the question-writing stage, at this point it may be desirable to commence *validity checking*. This is one of the verification elements of the process, which involves a panel of subject experts (usually university faculty, preferably independent). Their task is to address the validity of the instrument, in terms of *face validity* (i.e. does the diagnostic actually assess the skills it intends to assess) and *content validity* (i.e. does the diagnostic feature all the relevant elements of the topic area while excluding unrelated material). The suite of test objectives may be revised in light of the input from the expert panel.

The next stage is to write candidate test questions which address all the desired objectives. It is worthwhile to generate more than the minimum number needed for each objective, as some may need to be discarded later in the process due to reliability problems or to provide a balanced test. In selecting distracters, commonly-observed student misconceptions should be included. It may be useful to trial the questions in a free-response format (i.e. without multiple answer choices) with a small group of students; their answers can then be collated and any frequently-occurring errors or misconceptions adopted as distracters. In constructing the answer options, it is important to avoid what we have dubbed 'the Sesame Street effect' ("One of these things is not like the others"): none of the answers should stand out noticeably from the others due to length, style, or for language reasons. (One way to test for this is to give a trial set of students *only* the answer options, i.e. without the question, and check if any answers are unreasonably favoured.)

When a suitable bank of candidate questions has been generated, these can be assembled into an appropriately balanced prototype test, again with validity input from the expert panel. Questions and answers should be revised if necessary to enhance clarity and remove sources of ambiguity. Further trials with students, followed up with supporting interviews, are useful at this stage to make sure that students are interpreting the questions and answers in the manner intended by the setters (i.e. that the test has face validity).

The validated prototype test should now be ready for large-scale trials and verification of its *reliability*. The notion of reliability is quite distinct from validity: a reliable test is one which will give a consistent measure of students' competency in the relevant topics, and which will successfully discriminate between students with high and low ability. In essence, we seek to ensure that a student's test score is determined primarily by their actual facility in the targeted objectives, and not by some artefact of the test instrument, random chance, or some other external factor. Reliability verification is achieved by a statistical evaluation of the test responses of a large number of trial students. These should be drawn from as wide a sample as possible of the intended target population, i.e. from different classes, disciplines, institutions, years of study, and so on as appropriate. (Clearly, recruitment of colleagues from other departments/institutions for this trial phase is likely to be required, and should be set in motion early in proceedings.)

When the trial diagnostic instrument responses have been collated, the instrument reliability can be evaluated using a standard battery of statistical tests. In their description of the reliability verification for the Brief Electricity and Magnetism Assessment (BEMA), Ding et al. give a concise and lucid account of a set of five such appropriate statistical tests[8], which have become widely used for diagnostic test evaluation. We will not repeat their detailed treatment here, but qualitatively describe the statistical tests and refer the interested reader to their paper for mathematical details. The five reliability tests may be divided into two broad categories: those which focus on individual test items (but which nevertheless should also be examined from the perspective of the whole test), and those which assess the whole instrument as a unit. The former consist of the item *difficulty index* and *discrimination index* and the *point biserial coefficient*, whereas the latter are *Ferguson's delta* and the *reliability index*.

For each test item, the difficulty index is simply the ratio of the number of students who got the question correct to the total number of students who attempted the question. (Clearly, the more students who successfully complete the question, the higher the value of the difficulty index: for this reason, many suggest that it should more properly be called an 'easiness index'.) For a maximally discriminating diagnostic, a majority of questions with a difficulty index of about 0.5 is preferable,

though in practice this is clearly challenging to achieve and questions with difficulty indices in the range 0.3-0.9 are regarded as acceptable.

The discrimination index measures the extent to which a particular question successfully delineates between students with a firm grasp of the tested concepts and those with weaker knowledge. Questions with a high discrimination index strongly indicate whether a student getting them right is likely to do well overall. Conversely, any question with a negative discrimination index is more likely to be done correctly by the weaker rather than the stronger students; such a question is dysfunctional and should be amended or discarded. In general, a discrimination index of 0.3 or higher is desirable. The discrimination index can be calculated in two ways, either by dividing the trial cohort into two halves (with higher and lower overall scores) or by comparing the highest and lowest quartiles of the cohort. The second approach is more robust since with a normally-distributed cohort there will be a large number of students straddling the upper-half/lower-half boundary, but it does neglect half of the student responses so may be less desirable in cases where there is a limited volume of trial data.

The point biserial coefficient is a related concept to the discrimination index, and measures how strongly correlated the score of a single item is with overall scores on the complete test. Items with a high point biserial coefficient are consistent in performance with the remainder of the instrument. Consequently, questions with a low (or negative) coefficient feature student performance on a particular item which is not consistent with their performance on the test as a whole: such questions should therefore be considered for revision. A minimum value of 0.2 for the point biserial coefficient is the usual criterion.

The preceding three statistical tests are all applied to the individual questions making up the diagnostic instrument. Ideally, all questions should pass all the tests. However, a few outliers can be acceptable (particularly if there are compelling reasons for their presence, e.g. scene-setting questions or related, multi-part questions), provided that the values of these test statistics when averaged over the whole instrument lie within the recommended ranges.

In addition to these item-by-item statistics there are, as previously mentioned, two whole-test statistics to apply. Ferguson's delta is a test of discrimination, and measures how widely the scores of the trial student cohort are distributed over the possible range of test scores. An effectively discriminating diagnostic should have a large distinction between the scores of the stronger and the weaker students, and hence a broad range of overall scores, and consequently a large value of delta. Ferguson's delta values of 0.9 or above are generally considered acceptable.

The reliability index seeks to measure the repeatability, or self-consistency, of the test. Ideally, a reliable test given to the same student twice in quick succession should yield identical (or, at least, very similar) results. Clearly, actually doing so in practice is not feasible, not least of all because the student will remember the questions and their answers from the first iteration. The usual solution to this problem is to make use of a split-halves technique, in which the student's responses are divided into two halves, equivalent to them having completed two shorter tests in parallel: correlation between their scores on these half-tests can then be investigated. Clearly, this correlation will depend somewhat on exactly how the instrument is divided up. To address this, we may use Kuder-Richardson reliability formula 20 (KR-20), which averages over all possible combinations of half-tests. (For dichotomously scored tests such as those using MCQs, KR-20 is also exactly equivalent to Cronbach's alpha[26], another widely-used statistic.)

A related formula is Kuder-Richardson 21 (KR-21), which is simpler to calculate than KR-20 but makes the rather rigid assumption that all the test questions are of the same difficulty. KR-21 is reported in the BEMA reliability study and is also used elsewhere, e.g. in Wuttiprom *et al*.'s reliability study of the Quantum Physics Conceptual Survey[28]. However, in situations where this assumption is violated (which will almost always be the case with diagnostic tests) KR-21 will give only a lower bound on the true reliability[27], and may seriously underestimate the actual test reliability[29]. Since computer-based data processing techniques have become ubiquitous in the period since the introduction of the Kuder-Richardson formulae, there is now little additional burden in calculating KR-20 or Cronbach's alpha in preference to KR-21. For measuring the ability of groups of students (whole classes, etc.), the usual criterion is a value of the reliability index of 0.7 or higher.

In adopting split-halves measures such as these, there is an implicit assumption that all the test questions *should* be correlated with each other. This will be true if all test items are measuring a single 'construct' (which will indeed be the case for many diagnostic instruments), but if the diagnostic test in question is addressing more than one different (but presumably related) constructs then this assumption may not be valid. Thus, reliability indices such as these should be interpreted with caution.

It should be noted that the validity and reliability verification procedures evaluate the diagnostic instrument as a whole: for this reason, it is generally considered inadvisable to employ (or draw conclusions from) a limited subset of the test. If this is done, it should be done with care and with an eye to its limitations, as individual questions lack the robustness of the whole instrument and a restricted suite drawn from a larger test may not necessarily be reliable even if the whole instrument has satisfactory reliability.

Having evaluated the prototype diagnostic instrument using the statistical tests, any problematic items should be revised (or discarded if necessary). If the reasons for the dysfunctional nature of the questions are not clear, further triangulation via student interviews may be necessary, and if the modifications to the test have been substantial, further rounds of validation, trial deployment and reliability verification may be required. When a valid, reliable diagnostic instrument has been finalised, it can then be made available for widespread deployment. It is generally recommended that tests *not* be made freely accessible, either by publication or on the web, since their value will be very quickly compromised if students are able to see the content before testing. A common approach is to password-protect tests on the web, making the password available to instructors on request.

4. A brief survey of other tests in the literature

A plethora of other instruments have been developed and in this section we will give a brief tour through some of them. Links to many others can be found elsewhere[30]. We do not aim for a comprehensive list, but instead highlight certain families of instrument that have been created and validated and have subsequently been widely applied. Omissions are not through any lack of worthiness but principally due to lack of space.

Our first family tree to examine is the set of instruments that deal with the assessment of student expectations and beliefs about their subject (this area has been recently reviewed in a previous volume of this journal[31]). Elby has argued that ways of thinking about the subject that mirror those of expert practitioners – an 'epistemological sophistication' as he puts it - are extremely valuable, correlating with academic performance and conceptual understanding and supporting good study habits and metacognitive practices[32]. Such instruments - that aim to assess student attitudes and epistemologies - have a history of development back to William Perry's early work and have been active areas of development in the physical sciences almost as far back as the FCI (see, for example, Ref 33.) They include the Epistemological Beliefs Assessment for Physical Science (EBAPS)[34]; the Maryland Physics Expectation Survey (MPEX)[35], the Views About Science Survey (VASS)[36] and the more recent Colorado Learning Attitudes about Science Survey (CLASS)[37]. Our own investigations with the CLASS instrument have produced some interesting findings about the development of expert-like views as students proceed through their degree programme[38]. Similar findings have been reported from a comparable study at UCSD[39].

Another important family are the diagnostic tests that have been developed within the UK and deployed to provide quantitative evidence of the 'maths problem': the serious decline in mastery of the skills needed for mathematically based HE programmes. The key report on this, although now a decade old, is 'Measuring the Mathematics Problem'[40]. The serious (and rather bleak) picture painted at the time that report was written has improved somewhat over the last decade but many of the same challenges still remain for those teaching in introductory courses in mathematically-based disciplines. Several institutions have decades' worth of data on this topic but unfortunately rather little of it has found its way into journal or conference papers.

These sorts of assessment instruments add an important alternative perspective by not focusing on specific content knowledge. Related to this are assessment instruments that aim to appraise general attributes fostered and developed by a degree in the physical sciences. One of the most ubiquitous of such 'graduate attributes' in any science degree is the development of scientific reasoning and thinking skills. The Lawson Classroom Test of Scientific Thinking (LCTST)[41] was developed in the late 1970s and probes probabilistic thinking, identification of variables, proportional thinking and deductive reasoning via a series of 24 paired MCQ questions. Studies have found a strong positive correlation between students' normalised gains on the FCI and scores on the Lawson test[42]. A more recent study by Bao *et al.* has compared the scientific reasoning ability of post K-12 (final year) high school students in China and the USA and found broadly similar distributions of scores[43]. However, the same study showed that this similarity is starkly different to the same groups' performance on the FCI and BEMA instruments. Here, content knowledge and reasoning skills diverge, with the Chinese K-12 students significantly outperforming those from the USA.

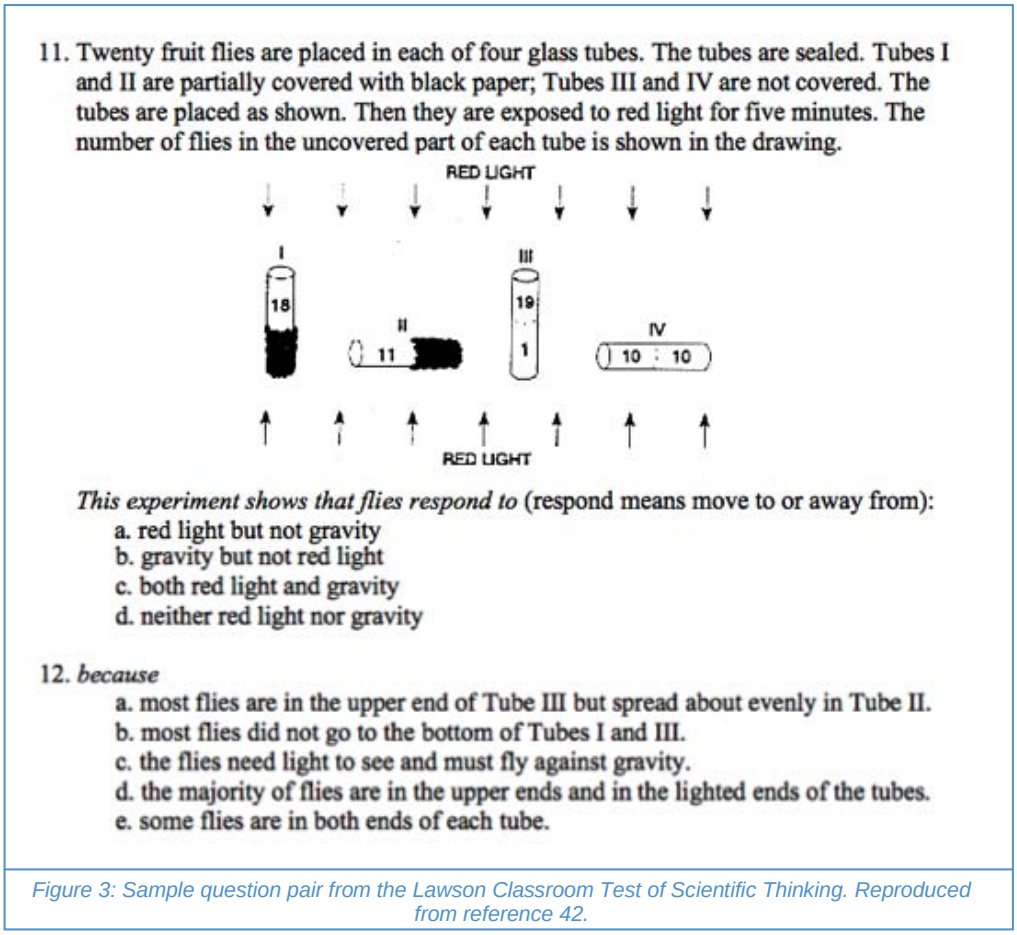

Figure 3: Sample question pair from the Lawson Classroom Test of Scientific Thinking. Reproduced from reference 42.

An example of a question from the Lawson test, reproduced from Coletta's paper[42], is shown in Figure 3 and illustrates the 'paired' nature of the questions on the test. As well as asking 'what', the second question of the pair asks for a 'why?' It is perfectly possible to have students reason what the correct answer is to the former, but choose one of the incorrect responses to the latter. In the case of this particular question pair, several of the 'why?' statements are valid, but do not fully

specify the correct reasoning. These sorts of two-tier questions in diagnostic instruments have been widely developed by David Treagust, looking at conceptual understanding in the field of chemistry (see for example[44,45]).

Our own studies in which we have used (parts of) the Lawson test, including the question pair illustrated in Figure 3, have revealed some interesting differences between student cohorts in physics on either side of the school-university transition. Whilst the percentages of each group that get the first part of the question pair correct show no statistically significant difference, around twice as many end-of-first-year undergraduate students choose the correct what-why pair of answers compared to students on the brink of entering university. This is statistically significant for the size of cohort groups we investigated (N = 80, 100, respectively).

In terms of challenging topics in a physical sciences degree programme, few can match quantum mechanics for its conceptual difficulty, counter-intuitiveness and a lack of real-world concrete experience. There has been much previous work in this area, including our own studies reported in a previous volume of New Directions[46]. The University of Colorado PER group have developed the Quantum Mechanical Conceptual Survey (QMCS), drawing on earlier work to develop a similar instrument. The test was devised using a two-tier free response approach, where an initial pilot version of the test asked students to identify 'what', followed by a free-text response area where they were asked to given a short reason 'why' they chose this. A detailed account of the construction and deployment of a test of conceptual understanding of introductory quantum mechanical concepts has been described by Wuttiprom *et al.*[28] This study covers not only details of the design and validation of the instrument, but assesses student performance and improvement after teaching on different types of questions (interpretive and non-interpretive).

It is no real surprise that these and other studies find compelling evidence of widely held alternative conceptions by students in the topic area of quantum mechanics, some of which persist after instruction. One of the most commonly-held alternate conceptions is to be found in the topic of quantum tunnelling. Students often remain convinced (even after instruction) that particles involved in tunnelling must lose energy, indicating a classical mental model of the process. This has been confirmed by our own investigation[46], the Colorado group[48] and a separate study from the University of Maine[49]. The latter study conducted a multi-year investigation using surveys, exams and interviews, and concluded that '*the response that 'particle energy is lost in tunneling' is prevalent across all our studies*'. This particular topic is one where visualisation is key to understanding: several groups have reported effective use of simulations in supporting and improving students' conceptual understanding of these topics[50,51,52]. Other approaches have also been described, such as vicarious learning through student-tutor discussions[53].

One final topic area in which we highlight instrument development is that of astronomy. Bardar *et al.* have reported the construction and validation process, together with field trials, of the Light and Spectroscopy Concept Inventory (LSCI)[54]. In this volume, Balfour and Kohnle present a fuller summary of available diagnostic tests in this area, together with results from their own instruments[55].

5. A case study: the Edinburgh Data Handling Diagnostic Instrument

To conclude our review, we illustrate some of the preceding points by way of a brief case study of one of our own diagnostic tests, the Edinburgh Data Handling Diagnostic (DHD). A full account of the development and validation of this instrument, along with its initial findings, will be given in a forthcoming publication[56].

The initial motivation for the creation of this instrument originated with a small-scale trial of a short (10 question) pre-prototype test of laboratory data analysis skills developed by a team at the University of British Columbia (UBC). Results from this trial were not fully conclusive (not least of all because it proved to be too difficult for undergraduate-level students), but nevertheless there was a strong suggestion that our undergraduate students were not developing the mastery of data handling skills that we might expect from the four or five years of a physical science degree.

Quantitative data handling skills, i.e. the ability to assess the quality of measured values, and process, display, interpret, and draw valid conclusions from them, constitute one of the most valued aspects of science degrees. Not only are they of key importance within the science disciplines themselves, but are also identified as being of essential utility by a wide variety of employers (see for example[57]).

However, these skills often form part of the so-called 'implicit curriculum' (see for example Atkinson's commentary[58]). The implicit curriculum is that set of learning outcomes which we all 'know' or 'expect' that out students will have acquired by the time they graduate, but which may not be explicitly taught or specifically assessed. In the case of practical data analysis skills, many degrees will feature specific tuition of the basics of data processing (means, standard deviations, etc.) in early years of the programme, but more advanced topics (non-linear model fitting, analysis statistics, etc.) are expected to be 'absorbed' as a side-effect of tackling in-depth experimental projects. Similarly, learning in the laboratory is often assessed by means of a standard 'lab report', which conflates many relevant but disparate elements, such as experimental competence, standards of record keeping, data processing, and clarity of expression. De-convolving a student's ability in one defined element of the experimental process – such as data handling skills – from a composite measure such as this is not always practical or reliable.

In response to this identified need and clearly delineated area of interest, we set out to develop a diagnostic instrument which would be tailored to the appropriate curriculum content and set at a level suitable for evaluating the skills of physical science students at various points in their degree course. In consultation with colleagues with responsibility for laboratory and data analysis instruction at various levels (and who formed part of our 'panel of experts'), we identified relevant topics for inclusion, incorporating such areas as accuracy & precision, functional forms, line fitting and quantitative error analysis.

The development team then generated a large bank of candidate questions, assembled a prototype instrument, and iterated it through multiple versions with validation input from both the expert panel and some trial students. When an instrument with appropriate balance had been produced – and

in which we could be confident there were no obvious sources of confusion or ambiguity – we proceeded to the large-scale trial phase of the development process. This involved trial deployment of the prototype 23-item instrument to over 1200 students in ten institutions across the UK and Ireland. The students were drawn from all educational levels (Scottish first year to final year Honours) and from both physics and chemistry departments. (This trial cohort was fairly large – many diagnostic instruments are reliability-tested with cohorts of a few hundred.)

Responses from the trial students were analysed en masse in order to assess the reliability of the instrument, using the statistical tests outlined in section 3. The diagnostic was found to perform satisfactorily. Four items required minor revision in light of their statistical performance and on feedback from trial students. (One of these was the first question, which proved statistically problematic since it was substantially too easy; however, it was retained as-is since it served an important purpose in scene-setting and as a confidence-builder, a factor which was rated as highly valuable in trial student feedback, particularly in light of the excessively difficult UBC pre-prototype.)

Satisfied that our diagnostic instrument was performing well, we were then able to investigate overall student performance on the test. Figure 4 shows an example of the response profile from a single test question. As may be seen, only a minority of students choose the correct answer (B). Of those answering incorrectly, the vast majority choose the same wrong answer (C). This is the hallmark of a classic misconception: if students have not been instructed on a topic or simply don't know how to do it (or are guessing), then the incorrect answers will be fairly evenly distributed amongst all the distracters (and indeed such answer profiles are seen for many of the DHD questions). A profile such as that seen in Figure 4 indicates something different: students do *think* they know how to do something, but are consistently doing it incorrectly (in this case confusing the standard error on the mean with the standard deviation). Response profiles such as this will either confirm previously-known alternate conceptions or highlight areas in which they also exist and to which more instruction should be focussed.

Comparison of the mean test scores between different classes was also informative. Our early suspicions about the development of data handling skills were confirmed: between the first and second years of our own degree programme there was a significant increase in ability, but thereafter (from second to fifth year) the mean test scores stagnated, with no statistically significant changes. A similar picture is seen nationally: for those institutions for which we have longitudinal data (i.e. test scores from more than a single year group), in only one instance was there a significant improvement in a later year of study (and again this was between a first and second year class).

You want to measure how long seals can stay under water by measuring the time between a dive and a resurface with a stopwatch. After watching a seal dive three times you have measurements of 13 minutes, 7 minutes and 10 minutes. What do you determine as the best estimate and standard error of these measurements?

Tip: In this case the standard deviation is found from: $\sigma = \sqrt{\frac{1}{N-1}\Sigma(x_{obs} - x_{ave})^2}$

A 10.0 ± 1.0 minutes

B 10.0 ± 1.7 minutes

C 10.0 ± 3.0 minutes

D 10.0 ± 5.2 minutes

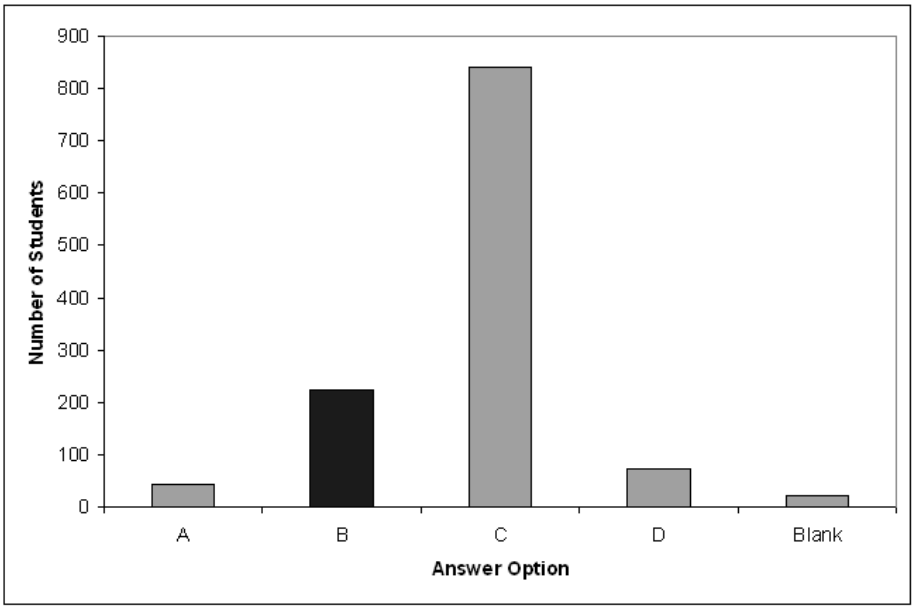

Figure 4: Sample question and student response profile (N ~1200) for one of the items in the Edinburgh Data Handling Diagnostic instrument. Item B is the correct answer.

In fact, the national picture is surprisingly uniform across all disciplines, institutions, and years of study: the variance of the mean scores of all the classes is less than all of the individual within-class variances, indicating that there is a relatively broad spread of individual student ability in data handling, but that on average most classes perform similarly, irrespective of discipline, location, or number of years of instruction. (Question-to-question success rates also show a striking correspondence, suggesting that the topics that the students do know are also fairly uniform.)

...a properly validated and robust diagnostic instrument ... affords an extremely powerful means to confirm the effectiveness of our teaching or to provide evidence-based guidance on which areas should be targeted for attention

The final phase of our project to develop the Data Handling Diagnostic is the generation of supporting learning resources, to be made available on-line. Students taking the test can be directed to these resources, which will provide additional instruction, explanation and practice for those areas in which they performed poorly. Follow-up steps such as these are a vital component of the diagnostic testing process: creation and deployment of a diagnostic instrument is only the first step, and must be followed with an analysis of class-aggregated and individual performance and such intervention as is found to be required (from extra supporting resources, through additional classes, to curriculum re-design if necessary).

As we have seen, as well as providing direct and useful feedback to students, a properly validated and robust diagnostic instrument (whether taken from the already wide-ranging literature or home-grown) affords an extremely powerful means to confirm the effectiveness of our teaching or to provide evidence-based guidance on which areas should be targeted for attention. Thus, diagnostic tests constitute one of the most valuable elements in the toolbox of the evidence-based educator.

Acknowledgements
Our understanding of and experience with diagnostic tests in the literature has been motivated and supported by Development Project funding in 2008/09 and 2009/10 from the Higher Education Academy UK Physical Sciences Centre. We also gratefully acknowledge support from colleagues in Physics and Astronomy at UBC (Doug Bonn and James Day) for providing the germ of the idea that led to the development of our DHD, and their permission to incorporate three of their pre-prototype questions in it.

References

1. MCQs have their detractors, but this is not the place to weigh the evidence for and against, nor to offer guidance on writing effective questions: plenty of such resources exist elsewhere, a few are collected in Ref. 2.
2. Wieman C.E. and Perkins K. (2005) *Transforming physics education*, Physics Today, November 2005, 36-41; Beatty I.D., Gerace W.J., Leonard W.J., Dufresne R.J. (2006) *Designing effective questions for classroom response system teaching*, Am. J. Phys., **74** (1), 31-39; Aubrecht G.J. and Aubtrecht J.D. (1983) *Constructing Objective Tests*, Am. J. Phys., **51** (7), 613-620; Bao L. and Redish E.F., (2000) *What can you learn from a (good) multiple choice exam?* [online] GIREP Conference Physics Teacher Education beyond 2000, Barcelona Spain, August 2000. <www2.physics.umd.edu/~redish/Papers/BRBarcelona.pdf> [accessed 21st May 2010].
3. Ausubel, D. P. (1968). Educational psychology: A cognitive view. New York: Holt, Rinehart & Winston.
4. Faculty of Science, The Open University, Are you ready for Science? (2010) [online] <www.open.ac.uk/science/courses-qualifications/are-you-ready-for-science/index.php> [accessed 29th May 2010].
5. Pollock S.J. (2009) *Longitudinal study of student conceptual understanding in electricity and magnetism*, Phys. Rev. Spec. Top. Physics Education Research, **5** (2), 020110.
6. Hake R.R. (1998) *Interactive-engagement Versus Traditional Methods: A Six-thousand-student Survey of Mechanics Test Data for Introductory Physics Courses*, Am. J. Phys., **66** (1), 64-74.
7. Hestenes D., Wells M., and Swackhamer G., (1992) *Force Concept Inventory*, Phys. Teach. **30** (3), 141-158; Halloun I. and Hestenes D., (1985) *The Initial Knowledge State of College Physics Students*, Am. J. Phys., **53** (11), 1043-1055; Halloun I. and Hestenes D., (1985) *Common Sense Concepts About Motion*, Am. J. Phys., **53** (11), 1056-1065.
8. Ding L., Chabay R., Sherwood B., Beichner R., (2006) *Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment*, Phys. Rev. Spec. Top. Physics Education Research, **2** (1), 10105.
9. We use the term ‘problems’ cautiously here: many of the questions will be more appropriately called ‘exercises’. A ‘problem’ can be defined as where the route to the solution is not immediately clear to the person trying to solve it.
10. Mazur E. (1997) Peer Instruction: A User's Manual Prentice-Hall, Upper Saddle River, NJ.

11. Savinainen A. and Scott P. (2002) *The Force Concept Inventory*, Physics Education, **37**, 45-52.
12. Alternative calculations of <*g*> have been proposed, based on different methods to calculate the quantity to account for the bias in low pre-test scores, negative gains for individual students and guessing. Ref. 13 contains a selection.
13. Bao L. (2006) *Theoretical comparisons of average normalized gain calculations*, Am. J. Phys., **74**, 917-922; Marx J.D. and Cummings K. (2007) *Normalized change*, Am. J. Phys. 75 87-91; Stewart J. and Stewart G. (2010) *Correcting the Normalized Gain for Guessing*, The Physics Teacher, **48**, 194-196.
14. Bloom B.S. (1984) *The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring*, Educational Researcher, **13** (6), 4-16.
15. Hake R.R. (1998) Interactive-engagement methods in introductory mechanics courses. [online] <www.physics.indiana.edu/~sdi/IEM-2b.pdf> [accessed 28th May 2010]; Hake R.R. (1999) Analyzing Change-Gain scores. [online] <www.physics.indiana.edu/~sdi/AnalyzingChange-Gain.pdf> [accessed 2nd June 2010].
16. Huffman D. and Heller P., (1995) *What Does the Force Concept Inventory Actually Measure?* Phys. Teach., **33** (3), 138-143; Hake, R.R. (2002) *Lessons from the physics education reform effort*, Conservation Ecology, **5** (2), 28 [online] <www.consecol.org/vol5/iss2/art28> [accessed 21st May 2010].
17. Mahajan S. (2005) Observations on teaching first-year physics. [online] <arxiv.org/abs/physics/0512158> [accessed 19th Sept 2006].
18. Hake R.R. (2001) Suggestions for Administering and Reporting Pre/Post Diagnostic Tests, [online] <www.physics.indiana.edu/~hake/TestingSuggestions051801.pdf> [accessed 31st May 2010].
19. Bates S.P. (2005) Reshaping Large Undergraduate Science Courses; the Weekly Workshop [online] CAL-laborate, UniServe Science International Newsletter, **14,** 1-6. [online] <science.uniserve.edu.au/pubs/callab/vol14/cal14_bates.pdf> [accessed 24th Nov 2007].
20. Bates S.P., Howie K. and Murphy A.S., (2006) *The use of electronic voting systems in large group lectures:challenges and opportunities*, New Directions: the Journal of the Higher Education Academy Physical Sciences Centre (ISSN 1740-9888), **2,** 1-8.
21. Hestenes D. and Wells M., (1992) *A mechanics baseline test*, Phys. Teach., **30** (3), 159–166.
22. Halloun I., (2010) Halloun Research on Modeling Theory, Profile Shaping Education and Educational Measurement – Assessment [online] <www.halloun.net/index.php> [accessed 3rd June 2010].
23. Thornton R.K. and Sokoloff D.R., (1998) *Assessing Student Learning of Newton's Laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula*, Am. J. Phys., **66** (4), 338-352.
24. Engelhardt P.V., (2009) [online] An Introduction to Classical Test Theory as Applied to Conceptual Multiple-choice Tests, in *Getting Started in PER*, edited by C. Henderson and K. A. Harper (American Association of Physics Teachers), Reviews in PER, **2**, <www.per-central.org/document/ServeFile.cfm?ID=8807> [accessed 28th May 2010]
25. Beichner R. J., (1994) *Testing Student Interpretation of Kinematics Graphs*, Am. J. Physics, **62**, 750.
26. Cronbach L.J., (1951) *Coefficient alpha and the internal structure of tests*, Psychometrika, **16** (3), 297-334.
27. Kuder G.F., and Richardson M.W., (1937) The theory of the estimation of test reliability, Psychometrika, **2** (3), 151-160.
28. 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.
29. Tucker L., (1949) *A note on the estimation of test reliability by the Kuder-Richardson formula 20*, Psychometrika, **14** (2), 117-119.
30. NC State University (2007) Assessment Instrument Information Page [online] <www.ncsu.edu/per/TestInfo.html> [accessed 21st May 2010].
31. El-Faragy N., (2009) *Epistemological beliefs and intellectual development in the physical sciences,* New Directions: the Journal of the Higher Education Academy Physical Sciences Centre (ISSN 1740-9888) **5**, 1-6.
32. Elby A., (2001) *Helping physics students learn how to learn*. Am. J. Phys., **69** (7), S55-S64.
33. Hammer, D., (1994) *Epistemological Beliefs in Introductory Physics*, Cognition and Instruction, **12** (2), 151-183.
34. Elby A., (1999) EBAPS [online] <www2.physics.umd.edu/~elby/EBAPS/home.htm#> [accessed 2nd June 2010].
35. Redish E.F., Steinberg R.N., and Saul J.M., (1998) *Student Expectations in Introductory Physics*, Am. J. Phys*.,* **66** (3), 212-224.
36. Halloun, I. and Hestenes D., (1998) *Interpreting VASS Dimensions and Profiles for Physics Students*, Science & Education, **7** (6), 553-577.
37. Adams W.K., Perkins K.K., Podolefsky N.S., Dubson M., Finkelstein N.D. and Wieman C.E. (2006) *New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey*, Phys. Rev. Spec. Top. Physics Education Research, **2** (1), 10101.
38. Bates S.P. and Slaughter K.A., (2009) *Mapping the transition - Content and pedagogy across the school-university boundary*, New Directions: the Journal of the Higher Education Academy Physical Sciences Centre (ISSN 1740-9888), **5**, 35-39; Bates S.P., Galloway R.K., Slaughter K.A. in preparation.
39. Gire E., Jones B. and Price E., (2009) Characterizing the epistemological development of physics majors. Phys. Rev. Spec. Top. Physics Education Research, **5** (1), 010103.
40. Hawkes T. and Savage M., (2000) [online] Measuring the Mathematics Problem <www.engc.org.uk/ecukdocuments/internet/document%20library/Measuring%20the%20Mathematic%20Problems.pdf> [accessed 19th May 2006].

41. Lawson A.E., (1978) *The development and validation of a classroom test of formal reasoning*, J. Res. Sci. Teach., **15** (1), 11-24: Lawson A.E. (2010) Anton Lawson assessments [online] <http://www.public.asu.edu/~anton1/LawsonAssessments.htm> [accessed 7th June 2010].
42. Coletta V.P. and Phillips J.A., (2005) *Interpreting FCI scores: Normalized gain, pre-instruction scores, and scientific reasoning ability*, Am. J. Phys., **73** (12), 1172-1182.
43. Bao L., Cai L., Koenig K., Fang K., Han J., Wang J., Liu Q., Ding L., Cui L., Luo Y., Wang Y., Li L., Wu N., (2009) *Learning and scientific reasoning*, Science, **323**, 586–587.
44. Peterson R.F. and Treagust D.F., (2006) *Development and application of a diagnostic instrument to evaluate grade-11 and-12 students' concepts of covalent bonding and structure*, Journal of Research in Science Teaching, **26** (4), 301-314.
45. Goh N.K., Tan K.C.D., Chai L.S. and Treagust D.F. (2002) *Development and application of a two-tier multiple choice diagnostic instrument to assess high school students' understanding of inorganic chemistry qualitative analysis*, Journal of research in Science Teaching, **39** (4), 283-301.
46. Archer R.S.K. and Bates S.P. (2009) Asking the right questions: *Developing diagnostic tests in undergraduate physics,* New Directions: the Journal of the Higher Education Academy Physical Sciences Centre (ISSN 1740-9888), **5**, 22-25.
47. McKagan S.B. and Wieman C.E., (2006) Exploring Student Understanding of Energy through the Quantum Mechanics Conceptual Survey [online] <arxiv.org/abs/physics/0608244> [accessed 8th March 2008]
48. McKagan S.B. Perkins K.K. and Wieman C.E., (2008) *A deeper look at student learning of quantum mechanics: The case of tunnelling*, Phys. Rev. Spec. Top. Physics Education Research, **4** (2), 020103.
49. Wittmann M.C., Morgan J.T., and Bao L., (2005) *Addressing student models of energy loss in quantum tunnelling*, European Journal of Physics, **26**, 939–950.
50. Adams W.K., Reid S., LeMaster R., McKagan S.B., Perkins K.K., Dubson M., and Wieman C.E., (2008) A study of educational simulations part 1 - engagement and learning. Journal of Interactive Learning Research, **19** (3), 397–419.
51. Adams W.K., Reid S., LeMaster R., McKagan S.B., Perkins K.K., Dubson M., and Wieman C.E., (2008) A study of educational simulations part 2 - interface design. Journal of Interactive Learning Research, **19** (4), 551–577.
52. McKagan S.B., Perkins K.K., Dubson M., Malley C., Reid S., LeMaster R., and Wieman C.E., (2008) *Developing and researching PHET simulations for teaching quantum mechanics*, Am. J. Phys., **76** (4), 406–417.
53. Muller D.A., Sharma M.D., Eklund J., and Reimann P., (2007) *Conceptual change through vicarious learning in an authentic physics setting*, Instructional Science, **35** (6), 519–533.
54. Bardar E.M., Prather E.R., Brecher K., Slater T.F., (2007) *Development and Validation of the Light and Spectroscopy Concept Inventory,* Astronomy Education Review, **5** (2), 103-113.
55. Balfour J. and Kohnle A., (2010) *Testing Conceptual Understanding in Introductory Astronomy*, New Directions: the Journal of the Higher Education Academy UK Physical Sciences Centre (ISSN 1740-9888), **6** , 21-24.
56. Bates S.P., Galloway R.K, Maynard-Casely H., Singer H., Slaughter K.A., in preparation.
57. Rees C., Forbes P. and Kubler B. (2006) Student employability profiles: A guide for higher education practitioners, The Higher Education Academy, [online] <www.heacademy.ac.uk/assets/York/documents/ourwork/tla/employability_enterprise/student_employability_profiles_apr07.pdf> [accessed 14th June 2010].
58. Atkinson G.F. (1981) *The implicit curriculum*, J. Chem. Educ., **58** (7), 560.

...diagnostic tests constitute one of the most valuable elements in the toolbox of the evidence-based educator.