Development of the Concise Data Processing Assessment

James Day and Doug Bonn*

Department of Physics and Astronomy, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z1

(Received 11 February 2011; published 29 June 2011)

The Concise Data Processing Assessment (CDPA) was developed to probe student abilities related to the nature of measurement and uncertainty and to handling data. The diagnostic is a ten question, multiple-choice test that can be used as both a pre-test and post-test. A key component of the development process was interviews with students, which were used to both uncover common modes of student thinking and validate item wording. To evaluate the reliability and discriminatory power of this diagnostic, we performed statistical tests focusing on both item analysis (item difficulty index, item discrimination index, and point-biserial coefficient) and on the entire test (test reliability and Ferguson's delta). Scores on the CDPA range from chance (for novices) to about 80% (for experts), indicating that it possesses good dynamic range. Overall, the results indicate that the CDPA is a reliable assessment tool for measuring targeted abilities in undergraduate physics students.

DOI: 10.1103/PhysRevSTPER.7.010114 PACS numbers: 01.40.Fk, 01.40.gf, 01.40.G−

I. INTRODUCTION

With contemporary physics education research efforts, there exists an increasing demand for strategies that reliably measure student comprehension and evaluate the success of instructional techniques. The ability to handle real data is considered by many to be the most important skill to be developed for the novice physicist, of all that is or could be taught in an introductory physics laboratory. We have developed a short diagnostic tool that addresses some of the differences between experts and novices in their ability to handle real data, differences identified in the literature and that our teaching experiences have shown exist with physics students' laboratory skills.

One broad class of skills indispensable to the physics student is aptitude with measurement and uncertainty. Research by Séré *et al.* [1] has shown that students typically do not understand the need to make several measurements, do not possess a critical insight into the notion of confidence intervals, cannot distinguish between random and systematic uncertainties, and hold the general notion that the more measurements one makes, the "better" the result is, without understanding the nature of what is meant by "better." Work by Leach *et al.* [2] has revealed that students commonly believe that perfect measurements can, in principle, be made (i.e., measurements without uncertainty), think that the arithmetic mean should always be used to obtain a final result from a set of data, and claim that the average is all that matters when comparing any two data sets. Many such published findings align well with our personal teaching experiences. For example, that students believe in the existence of a "true value" when a measurement is made [3] is consistent with our observation that students are often unable to weigh the relative importance of numbers that have differing uncertainty or recognize whether numbers with an associated uncertainty are in agreement with one another. These are just a few examples of where a clear distinction between expertlike versus novicelike thinking can be identified (others may be found in Refs. [4–6], and the references therein).

A second broad class of skills that is essential to any student planning to pursue the sciences is facility with data, graphs, and models. That is, they need to be able to move readily between numbers, functions, and graphical representations. Students frequently have difficulties in reading and interpreting graphs, as has been well documented for the case of kinematics graphs [7,8] and, more recently, with calculus graphs [9] for which the focus is conceptual understanding and graphical interpretation of a function and its derivative. Our observations suggest that such graphical literacy becomes significantly reduced when students are presented with anything more challenging than a linear-linear plot; currently, this is a topic for which the research is highly limited. Furthermore, we have anecdotally observed that our students struggle in recognizing basic functional forms in data sets whether they are presented numerically or graphically. (Related misconceptions about astronomical magnitudes, a power-law response formulaically represented by a logarithmic scale in an obscure base, have been reported [10].) Our students also persistently believe that any rapidly rising or falling function is "exponential" and even fail to recognize the difference between data of the form 2^x and x^2. They struggle even further when coefficients are added

*bonn@physics.ubc.ca

Published by the American Physical Society under the terms of the Creative Commons Attribution 3.0 License. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

to these functions, coefficients whose values must be estimated from data. This practical understanding of functions is another area for which the research is very limited.

This basic facility with data, including such useful tools as rescaling axes, semilog, and log-log plots, is of use to students no matter what their future path in the sciences may be, so we regard them as core goals of any first-year physics laboratory. These aspects of managing and interpreting real data are examples of key characteristics that distinguish experts from novices and are a class of numeracy skills essential to the sciences and increasingly valuable in a data-filled world. A numerate (and literate) population is a primary goal of modern society. Numeracy (and literacy) skills carry the means by which children are equipped for the education processes on which their location in the adult world will depend [11]. There are of course many other things that can be taught in a first-year laboratory, but the goals outlined above comprise a basic set of skills that is testable in the form of a multiple-choice concept inventory.

Motivated by earlier standardized, multiple-choice tests [8,9,12–20], which were designed to measure what physics students learn with respect to a given set of concepts, we have developed and validated the Concise Data Processing Assessment (CDPA). This assessment instrument is intended to probe mainly students' abilities in the two broad aspects of data handling discussed above. The CDPA itself serves not as an improvement over any of these previous diagnostics but rather as a complement to them: it tests skills that the other diagnostics do not consider. Tests of their understanding of uncertainty range from simple decisions over significant figures to critical judgements based on differing uncertainty in measured quantities. Their fluency in graphs, functions, and data is tested by asking them to identify models that best describe various data sets in tabular and graphical form. The CDPA is also intended to be used to compare the effectiveness of various classroom instructional approaches from one year to the next.

In this paper we report on the development of the CDPA and on the evidence collected in support of its validity and reliability. This survey can be used to probe students' mastery of some difficult concepts related to collecting and processing data. The design process, outlined in Sec. III, aligns with professional criteria that have been established for educational testing as well as the components of assessment identified as requisite in a recent National Research Council (NRC) study on assessment [21]. Crucially, we make use of student interviews to uncover the different ways in which students think and to make certain that students interpret the question as we intended. *Validity* pertains to the degree to which the test actually measures what it claims to measure, and is also the extent to which inferences, conclusions, and decisions made on the basis of test scores are appropriate and meaningful. Validity can be established by the combination of student interviews and a consensus of expert opinion and is discussed in Sec. IV. *Reliability* refers to whether a test is consistent within itself and across time, and can be measured using statistical calculations that focus both on individual items and on the test as a whole. The results of these statistical metrics are presented in Sec. V.

II. BACKGROUND

The CDPA was originally created for measuring mastery of data handling skills for first-year physics laboratory students who are either majoring in physics or are enrolled in a team-taught, academically rigorous science program (in which biology, chemistry, mathematics, and physics are presented in a unified and integrated format), called Phys 107/109 and Science One, respectively. The lecture component of the course associated with this laboratory covers typical first-year physics material, dealing with conservation laws, angular momentum of rigid bodies, simple harmonic motion, and wave phenomena, as well as concepts of probability and kinetic theory. It is intended for students planning to take higher-level courses in physics and astronomy.

It should be made explicit that our first-year physics labs have considerably different learning goals from the traditional first-year physics lab. This context is highlighted as a caveat regarding the types of students or settings for which the tool should be used—often, data handling learning goals are not addressed until more advanced labs. While there has been considerable research on the extent to which labs can contribute to students' conceptual understanding of physics (see section IV C. of Ref. [22]), the lab for which the CDPA was designed is not primarily motivated by the aim to teach particular physics concepts or to reinforce what is taught in lectures and tutorials. This arises from the concern that introductory physics labs frequently have a large number of learning goals including mastery of particular equipment, computer software, statistical methods, physics content, as well as all of the goals regarding data that were outlined in the last section, leading frequently to cognitive overload. To avoid such overload for students, this lab is focused on broadly applicable skills that will be of value no matter what their later academic path may be (such basic skills are equally important to those students planning on pursuing a career in the health sciences [23,24]) and that can uniquely be addressed in a laboratory setting. Physics concepts can be carefully woven into such a course, but the primary goal is to develop a practical mastery at handling measurements of any kind. This includes skill at acquiring data, understanding the nature of uncertainty in measurements, and developing statistical and graphical methods for evaluating the data.

The particular goals of the diagnostic are aimed at two broad classes of difficulties that we have observed in first-year students and that also persist in many students in higher-level laboratory courses. The first of these

is proper understanding of the uncertainty attached to a measurement and how it is used. Students' difficulties with this have been explored extensively by the University of Cape Town group's research on students' ideas about measurement and uncertainty; there, they examine deeply held student misconceptions about measurement and ways to change their thinking [25], and have also developed complementary diagnostic tools for that purpose [20]. They differentiate novice and expert thinking in this area as ''pointlike'' versus ''setlike'' thinking. The expert notion of setlike thinking views a measurement and its uncertainty as a distribution of possible outcomes from an experiment. Pointlike thinking ascribes importance to particular values of the measurement. This includes issues such as students' concern for searching for the ''right answer,'' or a fixation on verifying the first number that was measured, or ascribing special importance to a number that appears twice in a set of repeated measurements.

We add to this work on uncertainty a second substantial issue which is students' ability to build mathematical models that fit their measurements, and to derive meaning from the success or failure of those models; that is, how to fit functions to data and draw inferences from those fits. An expert seamlessly connects together numbers, mathematical functions, and various graphical representations of them. For instance, a value that quadruples when its independent variable is doubled is a sign of a quadratic power law. Exponentials rise or fall by a multiplicative factor when the independent variable changes by a fixed amount. Semilog and log-log plots can help elucidate these relationships, but many students struggle with the fluid conversion between numbers, functions, and graphs, which in experiments are a conversion between data, models, and their plots. This is akin to the problem students have moving back and forth between mathematical and text-based representation of problems. The high-level goal is for the students to discover that science is not simply a static body of concepts and mathematics, but is based on empirical observation and experimentation, and connecting those together involves these concepts of measured numbers, their uncertainty, and their graphical representation.

III. DEVELOPMENT

The development of the CDPA involved several sequential steps. The creation of such assessment tools provides methods to compare instruction across institutions and over time in a calibrated manner. This methodology for test construction is outlined by Adams and Wieman [26], who summarize the procedures recommended in ''Standards for Educational and Psychological Testing'' [27] and highlight key points in the National Research Council's 2001 study of assessment [21]. We have adhered to the recommendations that an assessment be founded upon three reciprocally connected elements: *cognition*, the facets of achievement that are to be evaluated, *observation*, the tasks employed to amass evidence about students achievement, and *interpretation*, the techniques used to analyze the evidence resulting from the tasks.

Briefly, we first established learning goals for the course (Sec. III A): identifying explicitly what the students should be capable of doing by the end of the term. With that, we created questions (Sec. III B) that directly related to our major, course-level learning goals. These questions were then presented to the students at the end of term, who were given up to 30 min to provide full written solution to these questions. Upon completion of the test, student volunteers were found with whom to conduct interviews (Sec. III C), to help us understand why and where student reasoning failed on various questions. Employing different types of research procedures (personal experience plus student test answers and interviews) was done very deliberately to obtain accurate insights into students' understanding. An array of perspectives is essential if one hopes to triangulate upon an authentic and balanced view, rather than one that may be biased by examining only from a single perspective. Having a better-grounded understanding of student thinking in hand, we drafted multiple-choice versions of the very same questions (Sec. III D). These were then given to the following year's students and, again, were followed by more student interviews (Sec. III E). The instrument was also presented to a number of experts to help determine test validity. Item and test statistics were then calculated to characterize the test and identify potentially problematic questions (Sec. III F).

Our primary goal was to arrive at a test of minimum length that would yield scores with the necessary degree of reliability and validity for the intended uses described above. Abridged tests need not necessarily lose validity to their longer counterparts (Bell and Lumsden [28] state ''that all tests could be reduced [in length] by more than 60% without appreciable decreases in validity'') and in fact the validity decreases as a test gets longer, typically after a small number (7–12) of items [29]. We also aimed to create an assessment that would not overtax our students' endurance and be useful beyond first-year, all the while recognizing that the cost of large dynamic range is paid for in resolution of the instrument.

A. Learning goals

Learning goals are statements which define, in operational terms, what the student should be able to do by the end of a course. In addition to helping direct the design of curriculum, learning goals can also guide teaching and aid learning by edifying what students are expected to master [30], as well as informing our evaluation methods by having made explicit what teachers care for their students to have learned. Well-developed learning goals are conducive to an evidence-based approach to education.

For the Phys 107/109 and Science One lab course, from which this project initiated, there are currently 42 distinct

learning goals. Examples of the learning goals from which the CDPA was created include, but are not limited to, being able to

(1) weigh the relative importance of numbers that have differing uncertainty

(2) judge whether or not a model fits a data set

(3) linearize exponential distributions, by using semilog plots, and power-law distributions, by using log-log plots and power-law scaling

(4) extract meaning from the slope and intercept of data which have been linearized.

All learning goals are similar to the examples above, and the lab course for which the CDPA was initially designed concentrates on broadly applicable data handling skills.

B. Questions

Having identified the skills that students should possess by the end of term, we drafted ten questions (see the appendix for the final versions) that directly related to our major course-level learning goals. Items 1, 2, 5, 6, 9, and 10 are related to students' understanding of the meaning and use of uncertainty in measurements, and items 3, 4, 7, and 8 are measures of their ability to relate functions, graphs, and numbers. The questions were administered to three sections of about 25 students each at the end of term, and the ~75 represent those students who attended their lab section during the final week of courses. Each section had an approximately equal mix of Phys 107/109 and Science One students, who were given 30 min to complete all ten questions and were asked not to use a calculator. Their performance was motivated by the promise that their score on this set of questions could only have an upward influence on their final lab grade (up to 1% bonus).

C. Student interviews

Student interviews should always be used when developing educational tests, and the value of the kind of information extracted from such interviews is stressed in the 2001 NRC report [21], which states that ''the methods used in cognitive science to design tasks, observe and analyze cognition, and draw inferences about what a person knows are applicable to many of the challenges of designing effective educational assessments.'' A general finding of physics education research is that students can perform well on sophisticated tasks while still having serious misunderstandings about the underlying concepts. Performing interviews with students can help to identify such occurrences. The principles behind and practice of the interviews used in the development of the CDPA are described in some detail by Adams and Wieman [26].

We performed interviews with ten student volunteers, after they completed the assessment, to discover what content and wording was appropriate for the test. The students chosen for interviews were all of the students willing to sit with the instructor and the researcher following completion of the assessment. As such, we did not control for student ability when selecting our volunteers. The interviewers, the course instructor (Bonn) and a researcher unknown to the students (Day), asked students to explain the answers they had given and to expand on what particular terms or concepts meant to them [31]. The instructor was involved in the earliest interviews before learning that it is standard practice to have all interviews done by a researcher outside of the course instruction. The interviews were semistructured and rather flexible, allowing for new questions to be raised as a result of what the student replied. Sometimes the concepts we asked about were related to specific assessment questions (e.g., how does one treat data with differing uncertainty?) and other times the concepts were introduced by the students (e.g., ''human error''). Because the researcher (Day) was unknown to the students, he played the role of a novice physicist, allowing for very simple questions (e.g., ''what do you mean when you use the word exponential?'' or ''what exactly is radioactive decay?'') to be interpreted earnestly by the student, thus providing excellent insight into the student's thinking. The instructor asked primarily the scripted questions (e.g., ''how did you calculate your average?''). We observed nothing during the interviews to lead us to believe that students were intimidated by the 2-on-1 interview style, consistent with the fact that all those who volunteered knew they would be speaking with two interviewers. Written notes were taken during these preliminary interviews, but they were not audio recorded. Upon their completion, the interviewers held a debriefing to make sense of everything that was said and done by the student. These interviews helped to confirm that we were measuring student thinking as intended, as well as allowing for us to catch any flaws in the questions that might allow students to misinterpret what was being asked.

D. Multiple-choice questions

Possessing a set of students' written answers and the insight gained from the interviews, the questions were transformed into multiple-choice format. Great care went into crafting multiple-choice options that accurately reflected student thinking. The distractor options presented to the students in the multiple-choice version of the assessment are representative examples of the most commonly identified failure modes that were displayed in student thinking. These options for all items are explained in the appendix.

These new multiple-choice questions were then administered as a pre-test to 145 first-year physics students in the Phys 107/109 and Science One stream. Students were given 20 min to complete all ten multiple-choice questions and were asked not to use a calculator. Students recorded their name and identification number, to allow for matched pre-test and post-test data at the end of the term. To motivate effort on the pre-test, it was explained that

although their scores would not count towards the course grade, their scores could be confidentially returned to them on request and would assist both themselves and their instructors to know the degree and type of effort required for them to meet the primary, course-level learning goals. The assessment itself was not returned to the students and no mention whatsoever was made that the same test would be given as a post-test.

The instrument was also presented to a number of experts, as outlined in Sec. IV, in order to collect evidence for the determination of test validity.

E. Question validation student interviews

We conducted further interviews with student volunteers to validate the multiple-choice questions, i.e., to determine whether the students were interpreting the questions and answer options as we intended. Twelve student interviews were conducted by a single interviewer. Each interview lasted about 1 h, during which the students were asked to rephrase in their own words the questions they thought were being asked of them. Notes were taken during the interview and an audio recording of each interview was collected. The audio recording was then reviewed by the interviewer immediately after having met with the student, during which more detailed notes were taken and conclusions about question validity were drawn. Three of the 12 audio recordings, selected at random, were listened to by a third party who was blinded to the conclusions of the interviewer, in order to help validate the conclusions of the interviewer. It was determined that the wording of the CDPA was clear, and that the questions were consistently interpreted as intended.

These same students were also asked to explain the answers they had provided. Through this process, malfunctioning items were identified and subsequently corrected. For example, the first item, concerning a weighted average, originally presented data (see Fig. 1 for the form of this question) which allowed students to select the correct option for the wrong reasons. The original question presented 90 ± 8 mL/s and 100 ± 2 mL/s as the independent measures from which an appropriate average was to be calculated. Students could succeed using either the properly weighted average of 99.4 mL/s or the ''overlap value'' of 98 mL/s with the given container volume of 900 mL to arrive at 9.1 s as their forced-choice answer. Fortuitously, this provided an example of pointlike thinking, in which students ascribed importance to the particular value at which the extremes of two uncertainties met one another, rather than thinking of each measurement as a distribution. As another example, the sixth item (see Fig. 6), concerning a straight line of best fit, graphically presented data printed sufficiently small that some students were not noticing the data point in the bottom right corner of the figures. Therefore, we were able to identify that some students were getting the wrong answer but for the wrong reasons.

F. Item analysis

The assessment was given again to the same students during the final week of classes with no prior notification, with bonus course credit awarded for post-test performance. As with the pre-test, the assessment itself was not returned to the students, although it was announced that instructors would be willing to discuss questions on the assessment with interested students in private.

We calculate various descriptive statistics from the students' postcourse scores, with the details presented in Sec. V. In particular, we found that item difficulty, item discrimination, and point biserial correlation provided particularly useful information; primarily, these statistics aid in describing how the questions on the assessment relate to one another and to the test as a whole. We further calculate Ferguson's delta, which serves as a measure of the discriminatory power of the test as a whole, and the Pearson product-moment correlation coefficient, which serves as a measure of the reliability of the test.

It should be made clear, however, that these statistics are often used for single factor summative assessments of individual students, so the more common interpretation of their values and acceptable cutoff ranges do not necessarily hold. The CDPA examines multiple different facets of thinking and learning, rather than a single construct, and certainly does not cover all of the course material. While the intended use of the CDPA is to measure how well students are thinking like experts, the goal is not simply to obtain a summative assessment of student learning; we also want to provide formative assessment of teaching. Indeed, we consider the results from the students' CDPA scores as a group more important than ranking individual students, which is fundamentally different from many assessments [26].

IV. TEST VALIDATION

Test validation is the procedure by which evidence is gathered to determine if the test items satisfactorily represent a concept domain and whether the test measures the properties that it proposes to measure. Face validity is an estimate of whether an assessment seems to measure certain criteria (without guaranteeing that it actually does), or the validity of a test at face value. In different terms, a test can be said to have face validity if it appears it will measure what it is supposed to measure. Content validation is the evaluation of the correctness of the items constituting the assessment.

In collecting evidence to determine test validity, a request for expert feedback was solicited and an online version of the CDPA was made available to all faculty members of the Department of Physics and Astronomy at University of British Columbia (UBC), a structure which allowed for responses to be submitted anonymously. Twelve faculty supplied their responses, providing us with evidence of both face and content validity. Four

faculty openly contacted us to provide specific critiques of the test items. One explicit example of this type of feedback was to avoid using the term ''radioactive decay'' (originally used in item 1, instead of water flow rates); the concern was that this is a term with which students tend to have some very bizarre interpretations as to what is actually meant, and that it is best to avoid using it whenever possible as it is difficult to know what students will think it means. Another explicit example was that we had inadvertently used the term ''standard deviation'' when we meant ''uncertainty in the mean'' in item 10. We also received feedback on the specific wording of many of our questions. Minor changes that resulted from this feedback included altering ''the data below shows'' to ''the data below show'' and ''model/function'' to ''algebraic expression.'' An example of a larger change was editing item 4 from ''the log-log plot below shows the natural logarithm of the power radiated by an object as a function of the natural logarithm of its temperature'' to ''the log-log graph below shows the natural logarithm of the power emitted E, measured in Watts (W), by an astronomical object as a function of the natural logarithm of its surface temperature T, measured in Kelvin (K).''

The CDPA was also administered to 11 graduate students registered in a graduate course on teaching techniques in physics and astronomy, and four additional graduate student volunteers (i.e., teaching assistants) familiar with the Phys 107/109 and Science One laboratory. Their responses were considered along with those submitted by the faculty in collecting lines of evidence of both face and content validity.

Construct validity is a measure of whether an assessment is able to successfully distinguish between populations. Analysis of CDPA scoring shows that it measures certain expected results (see Table I). For example, using a t-test for pairwise comparisons between levels and a Bonferroni correction to address the problem of multiple comparisons, there were statistically significant differences at the $0.05/4 = 0.0125$ level (equal variance not assumed) between the scores for first-year and fourth-year students: $t(110) = 2.85$, $p = 4 \times 10^{-3}$; the scores for fourth-year students and faculty: $t(26) = 7.51$, $p = 6 \times 10^{-14}$; and the scores for graduate students and faculty: $t(38) = 4.94$, $p = 8 \times 10^{-7}$. There was not a statistically significant difference between the scores for fourth-year and graduate students: $t(56) = 1.17$, $p = 0.24$; these two means certainly are different, but we cannot know whether the size of that difference is scientifically trivial or important without more data. However, there was also a statistically significant difference between the scores for first-year students and second-year students whose previous physics labs did not include data handling learning goals: $t(109) = 3.75$, $p = 3 \times 10^{-4}$. Collectively, these results suggest that the CDPA is able to distinguish between populations in the novice-to-expert spectrum of facility with data. Specifically, our results suggest that as physics students progress through their program, they become more expert-like in their data handling abilities. We also observe that a large portion of the improvement in scores happens during the first year (cf. first- and fourth-year UBC physics pre-test results of Table I with first-year UBC physics post-test results of Table II).

Of course, there exist limitations in the extent to which our conclusions generalize to broader populations. While the students used in this study should be largely representative of typical UBC physics students, our findings may

TABLE I. Pre-test results.

Sample	Number of students	Mean score	Standard deviation	Standard error
UBC Phys 107/109	71	2.56	1.56	0.22
UBC Science One	74	2.74	1.57	0.18
UBC Phys 101	254	2.28	1.34	0.08
U of Edinburgh first-year physics	249	2.13	1.38	0.09
Overall first-year students	647	2.30	1.43	0.06
UBC second-year students	59	3.03	1.52	0.20
UBC fourth-year students	75	4.73	2.26	0.26
UBC graduate students	32	5.31	2.36	0.42
UBC faculty	12	8.00	1.21	0.35

TABLE II. UBC post-test results.

Sample	Number of students	Mean score	Standard deviation	Standard error
Phys 107/109	122	3.89	1.98	0.18
Science One	133	3.93	2.00	0.17
Overall	255	3.91	1.99	0.12

not be consistent with all groups of students. As such, issues of potential internal and external threats to validity are briefly noted. Internal validity is the confidence that can be placed in the cause and effect relationship of a scientific study. As this work did not employ an experimental design which examined the effect of manipulating a single variable while observing the outcome, no threats to internal validity exist. Threats to external validity are factors which limit the extent to which the findings of a study can be applied to populations beyond that which was studied. Some inherent risk exists in generalizing from our sample of first-year students at UBC to the larger target populations of introductory physics students elsewhere in the world. The generalization of our conclusions is also limited by the extent to which the items approximate the actual learning objectives of the subjects in the target population. For this assessment, such concerns could relate to whether or not the items are deemed by the students to be authentic versus contrived. Lastly, the students in this study were not required to complete the assessment and received minimal extra credit for their willingness to answer questions. These same students might have performed differently if given the same questions in a higher stakes scenario.

V. TEST RELIABILITY

Test validity is requisite to test reliability; if a test is not valid, then its reliability is moot.

While there are psychometric tests that will provide reliability information, many of the commonly used statistical tests are specifically designed for assessments that measure a single construct or factor. One characteristic of the CDPA is that it evaluates thinking about multiple concepts, so the results of statistical measures must be interpreted accordingly.

Pre-test performance on CDPA does not vary much among different populations of first-year students, as shown in Table I, and pre-test scores average around 23%. The Phys 101 population was only measured in the pre-test stage, and are generally those students who are not considering physics as their major. Random guessing on the CDPA produces a score of 23.5%.

Table I also includes data from the University of Edinburgh, where the CDPA was administered in their flagship course in introductory physics. This class is large (nearly 250 students) and the student mix is inhomogeneous (half are aspiring physicists and half belong to other, mainly science, programs). The CDPA was administered at the University of Edinburgh because of an expressed interest from their physics education research group to get a measure of the data handling abilities of their students. Their data are presented as a demonstration of the apparently universal difficulties novice physicists have concerning the management and sense-making of data. A finer inspection of the University of Edinburgh data, although not presented here, reveals strikingly similar item difficulties, discriminations, etc. While only pre-test data were collected from the University of Edinburgh, we are further reassured of the CDPA's reliability, detailed below, upon comparison of their results to ours.

In this paper we use only postinstruction data, summarized in Table II, for test statistics as we are focusing on evaluation of the CDPA and not on a comparison of student pre-test and post-test performance. The post-test population deviates slightly from normality, with small positive skewness (0.45) and weak platykurticity (3.19).

A. Item difficulty index

The item difficulty index (P) is defined as the proportion of individual students in a sample that correctly answer the item. It is a measure of how difficult (or easy) a certain item is. Items with P values of 0.50 are taken as ideal (with the caveat that items are not highly intercorrelated), as they provide the highest levels of differentiation between individuals in a group. A low P value does not inevitably imply a malfunctioning item: a good item might be answered incorrectly by a majority of students if it addresses a deeply rooted misconception or difficulty in reasoning that has not yet been reversed by instruction. Also worth noting is that the P value depends on the particular population taking the test. As shown in Table III, the items on the CDPA cover a reasonable range in difficulty from about 0.2 to nearly 0.8. The averaged difficulty index value of all items, commonly

TABLE III. Item descriptive statistics for the CDPA.

CDPA item	Difficulty index, P	Discrimination index, D	Uncorrected point-biserial index	Corrected point-biserial index, r_{pb}
1	0.27	0.51	0.44	0.24
2	0.77	0.41	0.44	0.24
3	0.27	0.40	0.45	0.25
4	0.48	0.47	0.42	0.19
5	0.26	0.33	0.33	0.12
6	0.18	0.25	0.36	0.17
7	0.38	0.46	0.43	0.21
8	0.44	0.57	0.53	0.32
9	0.52	0.45	0.41	0.17
10	0.36	0.47	0.45	0.23

used as an indication of the test difficulty, is 0.39 for the CDPA.

B. Item discrimination index

The item discrimination index (D) is a measure of how well each item in a test distinguishes between more and less competent students. The higher the D value, the better the item discriminates.

An extreme group method is used to calculate D. To begin, two groups of students are created: an upper group consisting of those having the highest overall test scores and a lower group consisting of those having the lowest overall test scores. A reasonable population percentage to use in creating these extreme groups is the upper and lower 21% of the distribution, as this is the critical ratio that separates the tail from the mean of the standard normal distribution of response error [32]. The P value for each group is then determined and the difference between the two is taken, giving D. The possible range of D values is from -1 (where everyone in the lower group answers a question correctly and everyone in the upper group answers incorrectly) to 1 (vice versa). An item is typically considered to have good discrimination if $D \geq 0.3$ [33]. As shown in Table III, the items on the CDPA cover a reasonable range in discrimination from about 0.25 to nearly 0.6. The averaged discrimination index value is 0.43 for the CDPA, satisfying the commonly used criterion of $\bar{D} \geq 0.3$ [33]. Question 6 is the only item on the CDPA which has D below the arbitrary cutoff 0.3 and is therefore, at worst, only a weak discriminator. We have retained it because we feel it tests an important concept (of weighing the relative importance of data points that have differing uncertainty).

C. Item-to-total correlation

Another assessment of items related to their discrimination index is the (corrected) Pearson point-biserial correlation coefficient [34]. This metric, which probes how responses to an item relate to the total test score, is given by

$$r_{\mathrm{pb}} = \frac{\bar{Y}_1 - \bar{Y}}{\sigma_y}\sqrt{p_x/q_x},$$

where $\bar{Y}_1$ is the mean of the corrected total test scores for those whose dichotomous response was 1, $\bar{Y}$ is the mean of the corrected total test scores for the whole sample, σ_y is the standard deviation of all scores on the corrected total test, p_x is the proportion of individuals whose dichotomous response was 1, and q_x is the proportion of individuals whose dichotomous response was 0. The correction entails a total score which excludes the response to the item in question [35], as total scores which include the item in question will possess inauthentically greater correlation than total scores consisting only of other items in the test, especially when the assessment possesses relatively few items. Values of this metric range from -1 to 1, with a positive value meaning greater correlation between item and overall score. Values of $r_{\mathrm{pb}} \geq 0.2$ are considered desirable [36]. These data are shown in Table III.

The averaged corrected Pearson point-biserial correlation coefficient $\bar{r}_{\mathrm{pb}}$ for the CDPA is 0.21. This relatively low value is not too surprising considering that the CDPA was designed to test multiple abilities in as few questions as possible. Furthermore, a minimum critical Pearson point-biserial correlation coefficient has been defined [37] as one which is 2 standard deviations above zero, with the standard deviation calculated by

$$\sigma_r = \frac{1}{\sqrt{N-1}},$$

where N is the sample size. The minimum critical Pearson point-biserial correlation coefficient from our data is 0.125, a value met or exceeded by all of the items on the CDPA.

D. Ferguson's delta

A measure of whole-test discrimination is Ferguson's delta δ [38], which investigates how broadly the total scores of a sample are distributed over the possible range. Calculation of Ferguson's delta relies on the relationship between the overall test scores of any two students. Defined as the ratio of the observed number of relations of difference to the maximum number of such relations, it is given in simplified form as

$$\delta = \frac{n^2 - \sum_i^{k+1} f_i^2}{n^2 - n^2/(k+1)},$$

where f_i is the frequency obtaining a score value i, in a test of k items, administered to n individuals.

The possible range of values for Ferguson's delta are from 0 (no distribution) to 1 (a rectangular distribution); the normal distribution yields a δ of 0.93 and generally a good test should have a $\delta \geq 0.9$ [39]. Ferguson's delta for the CDPA is 0.94.

E. Test-retest reliability

Internal consistency coefficients can be used as measures of task variability, but since the goal of the CDPA includes probing multiple concepts with a minimum number of questions and as the CDPA was not designed to measure individual students, task variability is not a good reflection of the reliability of the instrument. Administering the test to two equivalent populations and obtaining a test-retest stability coefficient was the method we chose for measuring reliability of the CDPA.

A test-retest stability coefficient measures the consistency of test results if the same test could be given to the same population again under identical circumstances. Of course, this is impossible because it would require that giving it the first time does not have any impact on the test takers or that they have not changed in any other way

TABLE IV. Summary of statistical test results for the CDPA.

Test statistic	Range	Reasonable lower bound	CDPA value ($N = 255$)
Item difficulty index, P	$[0, 1]$	≥ 0.3	0.39
Item discrimination index, D	$[-1, 1]$	≥ 0.3	0.43
Point-biserial coefficient, r_{pb}	$[-1, 1]$	≥ 0.2	0.21
Ferguson's delta, δ	$[-1, 1]$	≥ 0.9	0.94
Test-retest stability (Pearson)	$[-1, 1]$	≥ 0.7	0.80

between the first and second administrations. However, when administering tests to large university courses, one has the ideal situation: the test can be administered again the following year in the same course. The population of students who enroll in a course is very similar from one year to the next, provided the university maintains constant admissions criteria. Support for this claim can be based on more than a decade of results for the incoming population of our first-year physics lab students at UBC, which have been remarkably stable (i.e., less than 5% variation) over that period. Each year's students will have received like preparation for the course, will possess like university experiences, and will be of like demographics.

When using the test-retest method, reliability may be estimated with the Pearson product-moment correlation coefficient between two administrations of the same measure. It is a measure of the correlation between two variables and may range from -1 (strong negative correlation) to 1 (strong positive correlation), and is widely used in the sciences to gauge the strength of linear dependence between two variables. The Pearson coefficient is usually used to correlate two measures of the same test subject; here we have used it to correlate two measures (item difficulties) of the same assessment (CDPA). Explicitly, we used the post-test data from 2009 ($N_{2009} = 118$) and 2010 ($N_{2010} = 137$) in our calculation of the Pearson coefficient, and paired the item difficulty of question 1 in 2009 with the item difficulty of question 1 in 2010, the item difficulty of question 2 in 2009 with the item difficulty of question 2 in 2010, and so on. We measure a reliability of 0.80; a reliability coefficient of 0.7 is usually regarded as a minimum for tests which are to be used with individuals [40].

It is worth being explicit, at this point, in our deliberate omission of reporting a value for the Cronbach's alpha, an internal consistency coefficient commonly quoted as a measure of reliability for many other physics education research concept tests. Cronbach's alpha is primarily useful for a single construct test, as it depends on both the correlation between questions and the number of questions; the CDPA, however, is not a single construct test. In fact, having a high correlation between items, which results in a higher value for Cronbach's alpha, means that these items are redundant. The way a formative assessment of instruction is typically administered puts a premium on minimizing the time required to complete the assessment and hence the number of questions. Therefore, a low Cronbach's alpha on an assessment of this type would be quite reasonable, and a high Cronbach's alpha on a formative assessment of instruction does not guarantee that the test will be more reliable for its intended use, and may be an indication that there are redundant questions that should be removed.

VI. SUMMARY

We have created a diagnostic instrument, called the Concise Data Processing Assessment (CDPA), that probes students' thinking related to the nature of handling data. Such skills include being able to appropriately weight measurement uncertainties when calculating simple statistics and/or in fitting straight lines to linear data, correctly propagate measurement precision through a simple calculation, extract a mathematical description from numerically and/or graphically represented data, and properly accounting for uncertainties arising from a digital probability distribution function. We have outlined our method for the development and validation of this formative assessment. Evidence for validity of the CDPA was collected through interviews with students and through expert review, and was demonstrated with descriptive statistics which showed that outcomes increased with level of expertise. The difficulty, discriminatory power, and reliability of the CDPA have all been considered, and the results of five different descriptive statistical tests are provided in Tables III and IV. These results indicate that our instrument is sufficiently reliable for the purposes of probing how well students actually handle data, as well as for comparing the effectiveness of various classroom instructional approaches from one year to the next. Our focus now shifts to the perpetual task (see, for example, Ref. [6]) of helping our students to score better on the CDPA.

ACKNOWLEDGMENTS

This work has been supported by the University of British Columbia through the Carl Wieman Science Education Initiative. We would like to thank Wendy Adams and Carl Wieman for their help at various stages of the development and validation of this assessment, and Simon Bates for providing to us the data from the University of Edinburgh.

APPENDIX: CONCISE DATA PROCESSING ASSESSMENT

A copy of the instrument, the Concise Data Processing Assessment, is included below for teachers and researchers to use however they see fit (see Figs. 1–10). As explanations are provided for the multiple-choice options of each question, we recommend that the real name of the instrument not be used with students. Some of the captions also include discussion about potential or perceived weaknesses to the question itself and scores should be considered as an upper bound on a student's facility with data. For details concerning conventions for calculating and reporting uncertainties, please refer to Ref. [41].

1. Student A measures the flow rate of water coming from a tap and reports it to be (90 ± 12) millilitres per second. Student B follows a different measurement procedure and reports the flow rate to be (110 ± 1) millilitres per second. How long would it take to fill a 1 litre container?

(a) 10.0 s

(b) 9.1 s

(c) 11.1 s

(d) 9.5 s

(e) 10.6 s

FIG. 1. Item 1 assesses whether or not students know that the measurement with the larger uncertainty should carry less weight in the answer to the question. A weighted mean could be used, but the difference in uncertainty is large enough that a student could also arrive at the correct answer by simply discarding the poorer measurement. The distractors include (a) use of the unweighted mean, (c) discarding the more precise measurement, (d) using a value of 105.5 mL/s, which is the value centered in the gap between $90 + 12$ mL/s and $110 - 1$ mL/s, and, (e) added to balance the options. The concern might be raised about whether students worry about systematic error, and in the method of student B, in particular. This concern reflects a particular kind of expert point of view, but is not one that we believe to be valid from the student perspective. (Over the course of our interviews, we have not encountered any evidence of a student worrying about systematic error for this question; that is not to imply that it does not happen, only that it does not appear to be common.)

2. An astronomer determines that the mass of the star Alpha Ceristinnae is (3.1 ± 0.1) times the mass of the sun. If the mass of the sun is $1.98892 \cdot 10^{30}$ kg, then what is the mass of the star Alpha Ceristinnae expressed in kg?

(a) $(6.165652 \pm 0.198892) \cdot 10^{30}$ kg

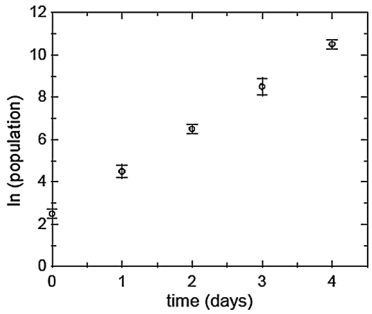
(b) $(6.16565 \pm 0.19889) \cdot 10^{30}$ kg

(c) $(6.2 \pm 0.2) \cdot 10^{30}$ kg

(d) $(6.166 \pm 0.199) \cdot 10^{30}$ kg

FIG. 2. Item 2 probes a student's judgment of an appropriate use of significant figures. The distractors include (a) the product with no rounding of significant figures, (b) the product rounded to the same precision as the multiplicand, and (d) the product rounded to an intermediate precision.

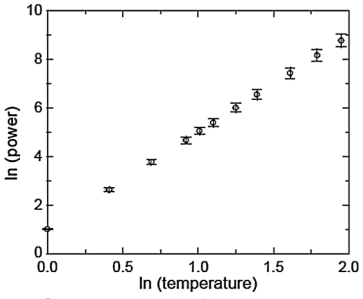
3. The semi-log graph below shows the natural logarithm of a population *N* of ocean water bacteria as it increases over time *t*. Which algebraic expression best describes this data?



(a) $N(t) = A \cdot t + B$ $\{A = 2\ \text{days}^{-1},\ B = 2.5\}$
(b) $N(t) = A \cdot e^{Bt} + C$ $\{A = 1,\ B = 2\ \text{days}^{-1},\ C = 12.2\}$
(c) $N(t) = A \cdot e^{t/B} + C$ $\{A = 12.2,\ B = 0.5\ \text{days},\ C = 0\}$
(d) $N(t) = A \cdot e^{Bt} + C$ $\{A = 12.2,\ B = 2\ \text{days},\ C = 0\}$

FIG. 3. Item 3 assesses a student's ability to interpret data displayed as a straight line on a semilog plot. The answer is most efficiently reached through applying knowledge that straight lines on semilog plots reflect exponential behavior, along with a basic knowledge of logarithm algebra and dimensional analysis. The distractors include (a) a straight line, ignoring the logarithm entirely, (b) the common problem of students attempting to work this out algebraically but failing to correctly manage the intercept in the calculation, and (d) the correct functional form but with incorrect units of the coefficient in the exponential. This question contains a hint to the correct answer in the form of physics content rather than pure knowledge of functions and graphs; that is, students might be helped by remembering that population growth is very often described by an exponential function.

4. The log-log graph below shows the natural logarithm of the power emitted *E*, measured in Watts (W), by an astronomical object as a function of the natural logarithm of its surface temperature *T*, measured in Kelvin (K). Which algebraic expression best describes this data?



(a) $E(T) = A \cdot T^B + C$ $\{A = 1\ \text{W/K}^4,\ B = 4,\ C = 2.7\ \text{W}\}$
(b) $E(T) = A \cdot T^B + C$ $\{A = 4\ \text{W/K},\ B = 1,\ C = 1\ \text{W}\}$
(c) $E(T) = A \cdot T^B + C$ $\{A = 2.7\ \text{W/K}^4,\ B = 4,\ C = 0\ \text{W}\}$
(d) $E(T) = A \cdot T^B + C$ $\{A = 2.7\ \text{W},\ B = 4\ \text{K},\ C = 0\ \text{W}\}$

FIG. 4. Item 4 assesses a student's ability to interpret data displayed as a straight line on a log-log plot. The answer is most efficiently reached through applying knowledge that straight lines on log-log plots reflect power-law behavior (with the power given by the slope), along with a basic knowledge of logarithm algebra and dimensional analysis. The distractors include (a) the common problem of treating the intercept on the graph incorrectly in their algebraic conversion of the logarithm, (b) in which the logarithms are ignored entirely and the formula for a straight line is given, and (d) the correct functional form but with incorrect units for the coefficients. This question contains a hint to the correct answer in the form of physics content rather than pure knowledge of functions and graphs; that is, students might be helped if they are cognizant of the Stefan-Boltzmann law.

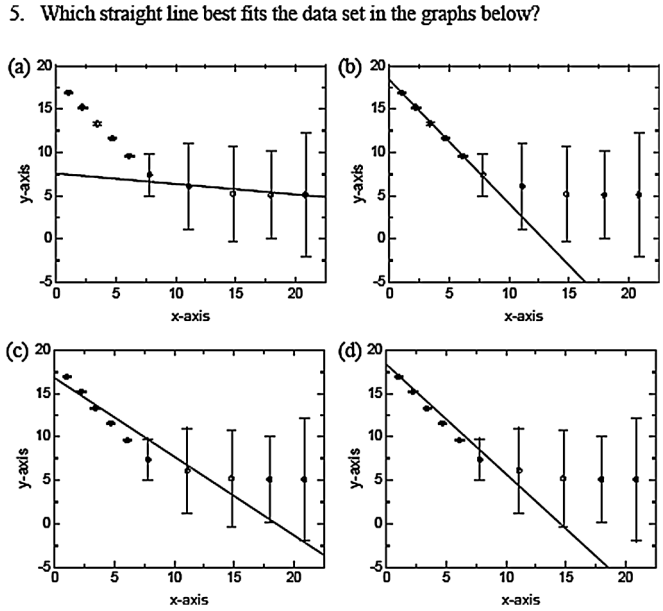


FIG. 5. Item 5 tests the student's ability to interpret uncertainties displayed graphically and how they should be employed in weighting a straight line fit. The distractors include (a) the problem where students ascribe greater importance to the prominence of the data with larger error bars, (c) comes from striving to have an equal number of points above and below the line as well as touching the error bars of the least important, but most prominent, data, and (d) which is close to a correct weighting, but still gives too much weight to the poorest data. A chi-square statistic of the numerical data clearly shows that fit (b) is best. These data might seem contrived but can actually be produced under very natural circumstances. For example, a semilog plot of the voltage decay across a capacitor in a *RC* circuit [i.e., $\ln(V)$ versus t] results in data and error bars of this sort.

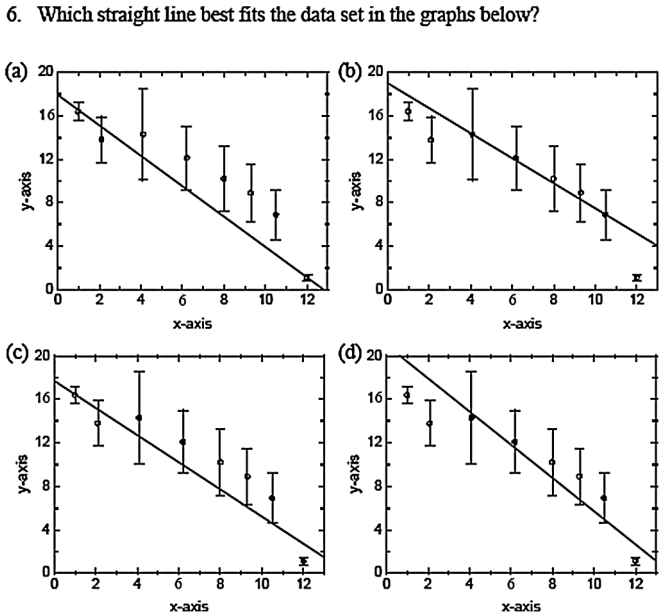
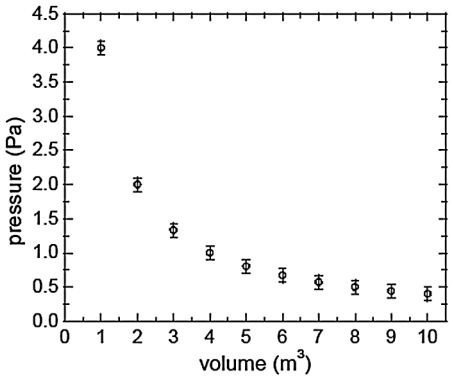


FIG. 6. Item 6 is a further test of interpretation of graphically represented uncertainty and data fitting. The distractors include (b) which appears to hit two points ''exactly'' and balances three points each above and below the line, as well as coming very close to four prominent points arranged close to a straight line, (c) which touches the maximum number of data points as possible within their error bars, and, (d) which possesses an equal number of data points above and below the line. A chi-square statistic of the numerical data clearly shows that fit (a) is best. It is possible that students select the correct answer to this question for the wrong reason, i.e., because they believe a line of best fit should pass through the first and last data points in a data set (what the University of Cape Town group would likely classify as ''point reasoning''); however, we have not encountered any evidence with our students that those (very few) who select the correct answer for this question are doing it for incorrect reasons. To eliminate this weakness, a next iteration of this question should make the second to last point have the extremal minimum uncertainty.

7. The graph below shows the relationship between the pressure P of a toxic gas at 20°C and the volume V of the safety vessel in which it is contained. Which algebraic expression best describes this data?



(a) $P(V) = A \cdot e^{B \cdot V}$ $\{A = 4\ \text{Pa/m}^3,\ B = -4\}$
(b) $P(V) = A \cdot e^{B \cdot V}$ $\{A = 4\ \text{Pa},\ B = -4\ \text{m}^{-3}\}$
(c) $P(V) = A \cdot V^B$ $\{A = 4\ \text{Pa}\cdot\text{m}^3,\ B = -1\}$
(d) $P(V) = A \cdot V^B$ $\{A = 4\ \text{Pa},\ B = -1\ \text{m}^{-3}\}$

FIG. 7. Item 7 tests the student's ability to identify a power law in graphically displayed data. The distractor options include the tendency to think that any rapidly falling function is exponential decay, with (a) incorrect and (b) correct dimensional units in the exponential, and (d) which gives the correct power law but mishandles the units. This question contains a hint to the correct answer in the form of physics content rather than pure knowledge of functions and graphs; that is, students might be helped by recalling the ideal gas law.

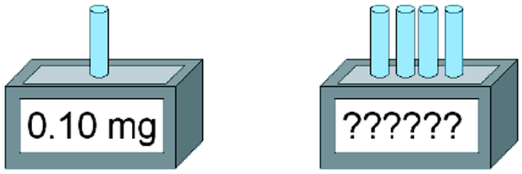
8. The data below show the measured luminous flux F of an incandescent light bulb at varying distances d separating the light bulb from the luminous flux detector. Which algebraic expression best describes this data?

distance (m)	measured flux (lm)
1	1.998 ± 0.002
2	0.501 ± 0.001
3	0.223 ± 0.002
4	0.125 ± 0.001
5	0.079 ± 0.003
6	0.056 ± 0.003
7	0.039 ± 0.004
8	0.030 ± 0.003
9	0.023 ± 0.005
10	0.018 ± 0.005

(a) $F(d) = A \cdot d^B$ $\{A = 2\ \text{lm}\cdot\text{m}^2,\ B = -2\}$
(b) $F(d) = A \cdot d^B$ $\{A = 2\ \text{lm/m}^2,\ B = -2\}$
(c) $F(d) = A \cdot d^B$ $\{A = 2\ \text{lm},\ B = -2\ \text{m}^{-2}\}$
(d) $F(d) = A \cdot e^{d \cdot B}$ $\{A = 2\ \text{lm},\ B = -2\ \text{m}^{-2}\}$

FIG. 8. Item 8 probes the student's ability to identify a power law in a column of numbers. The distractor options include (b) and (c) which both present the correct power law but mishandle the units, and, (d) which again checks on the tendency to default to exponential decay. This question contains a hint to the correct answer in the form of physics content rather than pure knowledge of functions and graphs; that is, students might be helped by remembering that the inverse-square law generally applies when some conserved quantity is radiated outward radially from a point source.

9. Imagine that you have used a digital weighing scale to measure the mass of a precisely machined stainless steel rod and the digital scale shows a reading of 0.10 mg. If you then measure a total of four of the same precisely machined rods on the scale, what are the possible values that the digital scale might read?



(a) 0.38, 0.39, 0.40, 0.41, or 0.42 mg
(b) 0.39, 0.40, or 0.41 mg
(c) 0.36, 0.37, 0.38, 0.39, 0.40, 0.41, 0.42, 0.43, or 0.44 mg
(d) 0.40 mg
(e) 0.40 or 0.41 mg

FIG. 9 (color online). Item 9 assesses both the understanding of rounding error in a digital instrument and the propagation of that error through a simple multiplication. The distractor options include (b) which arises from ignoring the uncertainty in the mass of a single rod, and then attaching an uncertainty of 0.01 mg to the meter, (c) attaches a 0.01 mg uncertainty to the initial measurement, rather than a 0.005 mg rounding uncertainty (note that some experts will do this since real meters often have instrumental uncertainty greater than the rounding error), (d) ignores uncertainty entirely, and, (e) which does not treat the rounding uncertainty symmetrically.

10. A small scale is used to measure the mass of several pieces of gravel and the mean mass is found to be 20.0 g, with an uncertainty in the mean of 4.0 g. On another scale, the mass of a small pile of gravel is found to be 6000 ± 20 g. What is a suitable estimate of the number of pieces of gravel in the pile?

a) 300 ± 1
b) 300 ± 5
c) 300 ± 24
d) 300 ± 60
e) 300 ± 80

FIG. 10. Item 10 assesses the student's ability to propagate uncertainty through a simple division, either by knowing the appropriate technique for handling multiple uncertainties or with a judgement that one of the uncertainties is completely dominant. The distractor options include (a) which arises from dividing the uncertainty in the numerator by the measured quantity in the denominator, (b) which arises from dividing the uncertainty in the numerator by the uncertainty in the denominator, (c) which arises from addition of the uncertainties, and (e) which arises from multiplication of the uncertainties.

[1] M.-G. Séré, R. Journaux, and C. Larcher, Learning the statistical analysis of measurement errors, Int. J. Sci. Educ. **15**, 427 (1993).

[2] J. Leach, R. Millar, J. Ryder, M.-G. Séré, D. Hammelev, H. Niedderer, and V. Tselfes, Survey 2: Students' images of science as they relate to labwork learning. Working paper 4, Labwork in Science Education. European Commission: Targeted Socio-Economic Research Programme, Project PL 95-2005 (1998).

[3] S. M. Coelho and M.-G. Séré, Pupils' reasoning in practice during hands-on activities in the measurement phase, Res. Sci. Technol. Educ. **16**, 79 (1998).

[4] D. L. Deardorff, Introductory Physics Students' Treatment of Measurement Uncertainty, Ph.D. thesis, North Carolina State University, 2001.

[5] R. F. Lippmann, Students' Understanding of Measurement and Uncertainty in the Physics Laboratory: Social Construction, Underlying Concepts, and Quantitative Analysis, Ph.D. thesis, University of Maryland, 2003.

[6] R. Lippmann Kung and C. Linder, University students' ideas about data processing and data comparison in a physics laboratory course, NorDiNa **4**, 40 (2006).

[7] L. C. McDermott, M. L. Rosenquist, and E. H. van Zee, Student difficulties in connecting graphs and physics: Examples from kinematics, Am. J. Phys. **55**, 503 (1987).

[8] R. J. Beichner, Testing student interpretations of kinematics graphs, Am. J. Phys. **62**, 750 (1994).

[9] N. Perez-Goytia, A. Dominguez, and G. Zavala, Understanding and interpreting calculus graphs: Refining an instrument, AIP Conf. Proc. **1289**, 249 (2010).

[10] E. Schulman and C. V. Cox, Misconceptions about astronomical magnitudes, Am. J. Phys. **65**, 1003 (1997).

[11] J. Bynner, Literacy, numeracy and employability: Evidence from the British birth cohort studies, Lit. Num. Stud. **13**, 31 (2004).

[12] R. K. Thronton and D. R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning and Lecture Curricula, Am. J. Phys. **66**, 338 (1998).

[13] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).

[14] P. V. Engelhardt and R. J. Beichner, Students' understanding of direct current resistive electrical circuits, Am. J. Phys. **72**, 98 (2004).

[15] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).

[16] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 010105 (2006).

[17] C. Singh and D. Rosengrant, Multiple-choice test of energy and momentum concepts, Am. J. Phys. **71**, 607 (2003).

[18] S. B. McKagan and C. Wieman, Exploring student understanding of energy through the Quantum Mechanics Conceptual Survey, AIP Conf. Proc. **818**, 65 (2006).

[19] S. B. McKagan, K. K. Perkins, and C. E. Wieman, Design and validation of the Quantum Mechanics Conceptual Survey, Phys. Rev. ST Phys. Educ. Res. **6**, 020121 (2010).

[20] www.phy.uct.ac.za/people/buffler/edutools.html.

[21] National Research Council, *Knowing What Students Know: The Science and Design of Educational Assessment*, edited by the Committee on the Foundations of assessment, J. W. Pellegrino, N. Chudowsky, R. Glaser, and Board on Testing and Assessment Center for Education Division of Behavioral and Social Sciences and Education (National Academy Press, Washington, DC, 2001).

[22] L. C. McDermott and E. F. Redish, Resource Letter: PER-1: Physics Education Research, Am. J. Phys. **67**, 755 (1999).

[23] S. L. Sheridan and M. Pignone, Numeracy and the medical student's ability to interpret data, Effect. Clin. Pract. **5**, 35 (2002).

[24] L. M. Schwartz, S. Woloshin, W. C. Black, and H. G. Welch, The role of numeracy in understanding the benefit of screening mammography, Ann. Intern. Med. **127**, 966 (1997).

[25] A. Buffler, S. Allie, F. Lubben, and B. Campbell, The development of first year physics students' ideas about measurement in terms of point and set paradigms, Int. J. Sci. Educ. **23**, 1137 (2001).

[26] W. K. Adams and C. E. Wieman, Development and validation of instruments to measure learning of expert-like thinking, Int. J. Sci. Educ. **33**, 1289 (2011).

[27] American Educational Research Association, American Psychological Association, and the National Council on Measurement and Education, *Standards for Educational and Psychological Testing* (American Educational Research Association, Washington, DC, 1999).

[28] R. Bell and J. Lumsden, Test length and validity, Appl. Psychol. Meas. **4**, 165 (1980).

[29] M. Burisch, Test length and validity revisited, Eur. J. Pers. **11**, 303 (1997).

[30] B. Simon and J. Taylor, What is the Value of Course-Specific Learning Goals?, J. Coll. Sci. Teach. **39**, 52 (2009).

[31] Typically, having the course instructor interview their own students is avoided. At the time these interviews were conducted we simply did not know any better.

[32] R. B. D'Agostino and E. E. Cureton, The 27 Percent Rule Revisited, Educational and Psychological Measurement **35**, 47 (1975).

[33] D. Doran, *Basic Measurement and Evaluation of Science Instruction* (NSTA, Washington, DC, 1980), p. 99.

[34] E. E. Ghiselli, J. P. Campbell, and S. Zedeck, *Measurement Theory for the Behavioral Sciences* (W. H. Freeman and Company, San Francisco, CA, 1981), p. 116.

[35] M. J. Allen and W. M. Yen, *Introduction to Measurement Theory* (Waveland Press, Long Grove, IL, 1979), p. 123.

[36] P. Kline, *A Handbook of Test Construction: Introduction to Psychometric Design* (Methuen, London, 1986), p. 143.

[37] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Holt, New York, 1986), p. 34.

[38] G. A. Ferguson, On the theory of test discrimination, Psychometrika **14**, 61 (1949).

[39] P. Kline, *Handbook of Psychological Testing* (Routledge, London, 2000), p. 31, 2nd ed.
[40] P. Kline, *The New Psychometrics: Science, Psychology and Measurement* (Routledge, London, 1998), p. 29.
[41] It is important to recall that different research areas use different conventions for reporting uncertainty. In the CDPA, we use $x \pm \Delta x$ to represent a mean (x) and an uncertainty in the mean (Δx). Interested readers might find ''Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results'' especially helpful, http://physics.nist.gov/Pubs/guidelines/.