Assessing the quality of a student-generated question repository

Assessing the quality of a student-generated question repository

Simon P. Bates*

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

Ross K. Galloway, Jonathan Riise, and Danny Homer

Physics Education Research Group, School of Physics and Astronomy, University of Edinburgh, Edinburgh EH9 3JZ, United Kingdom

(Received 20 February 2013; revised manuscript received 5 June 2014; published 18 July 2014)

We present results from a study that categorizes and assesses the quality of questions and explanations authored by students in question repositories produced as part of the summative assessment in introductory physics courses over two academic sessions. Mapping question quality onto the levels in the cognitive domain of Bloom's taxonomy, we find that students produce questions of high quality. More than three-quarters of questions fall into categories beyond simple recall, in contrast to similar studies of student-authored content in different subject domains. Similarly, the quality of student-authored explanations for questions was also high, with approximately 60% of all explanations classified as being of high or outstanding quality. Overall, 75% of questions met combined quality criteria, which we hypothesize is due in part to the in-class scaffolding activities that we provided for students ahead of requiring them to author questions. This work presents the first systematic investigation into the quality of student produced assessment material in an introductory physics context, and thus complements and extends related studies in other disciplines.

DOI: 10.1103/PhysRevSTPER.10.020105 PACS numbers: 01.40.Fk, 01.40.G-, 01.40.gb

I. INTRODUCTION

It has been argued that there are specific and defined educational benefits from students being engaged in the cocreation as well as consumption of educational content [1]. Cognitively, it can be far more challenging to have to create an assessment activity, for example, a question complete with solution and explanation, than it is to simply answer one created by someone else. It can require higher order skills far above simply "remembering'" or "knowing" in a facile sense, a process which will be familiar to many faculty as they regularly take up the challenge of setting end-of-course assessments that meaningfully assess learning goals. Others have framed the benefits associated with student-generated content in terms of a participatory learning approach [2] designed to foster deep, as opposed to surface, learning. A basic premise then is that by requiring students to develop assessment content themselves, we are challenging them to operate at higher cognitive levels than they might otherwise do in the normal course of their studies, encouraging a deeper approach to learning.

Skills in the cognitive domain of Bloom's taxonomy [3] provide a useful framework for categorizing the cognitive level or degree of challenge associated with learning materials. Ascending the taxonomy levels (which are often represented diagrammatically as a pyramid structure) are descriptors of knowledge—understanding-application-analysis-synthesis-evaluation. Anderson and Krathwohl [4] have suggested a simplification or revision to this structure that brackets the uppermost three levels of analysis-synthesis-evaluation together into a single compound category. A similar categorization, utilizing the same taxonomy, has been proposed as an aid to creating and refining the learning outcomes or goals associated with courses [5].

This article describes the incorporation of student-generated content in the context of multiple choice assessment questions (MCQs) in introductory physics courses delivered at a large, research-intensive university in the UK. Specifically, we investigate the quality of questions that students authored as part of their assessed coursework, by mapping a representative fraction of these onto the levels of Bloom's taxonomy. Additionally, we characterize the quality of student-produced explanations associated with these questions. Though there are some systematic investigations into the quality of student-produced materials reported in the literature (which we briefly discuss in the following section), we are not aware of any within the physics education literature, a gap which the present study attempts to address.

II. BACKGROUND AND CONTEXT

One of the key features of the coming of age of modern information technologies (for example, within the so-called

*simon.bates@ubc.ca

"Web2.0" movement) has been a shift from a single, authoritative content owner, dispensing knowledge and information to those who consume it, towards a much more collaborative approach, with potentially large numbers of coproducers of content. An often-cited example is the development of Wikipedia which, despite some concerns over the quality of some content, continues to be one of the ten most frequently accessed websites across the world [6], with a team of committed content authors well in excess of 300 000 and over 18 million occasional contributors [7].

Within the domain of education, earlier studies in psychology have shown that the act of question writing can be an effective study and learning technique, with question authors outperforming nonauthoring students on subsequent tests, irrespective of whether they wrote essay-type or multiple choice questions [8]. Multiple choice questions are often viewed as quite limited in terms of their assessment potential, particularly when their use is driven by staff needs for greater efficiencies, rather than student needs for deeper understanding. However, it has been shown that they can support the process of formative assessment and feedback, as described by the "seven principles" of good practice identified by Nicol and Macfarlane-Dick [9]. A key finding from a related review [10], which focused on effective e-assessment by design using MCQs, highlighted the importance of not just the questions themselves, but the context in which they were deployed within courses. A case study within that review presents student creation of assessment questions as a powerful articulation of Nicol and Macfarlane-Dick's first principle, that students understand what is required for good performance in terms of goals, criteria, and standards.

Along with instructors at more than seven hundred institutions worldwide, we have been using the PeerWise online tool [11] as the technology platform for these interventions. Developed in the department of computer science at the University of Auckland, PeerWise is a freely available, online tool to facilitate cohorts of students writing their own MCQs and answering and commenting on those of their peers. It incorporates much of the social functionality found in common popular websites, such as the ability to rate and comment on posts, and "follow" other contributors. At the time of writing, nearly 100 000 student registrants have contributed around 600 000 questions and approximately 12 million answers. In the last few years, several studies have emerged that have assessed the impact on student engagement and learning [12–21] including our own initial study [22] of piloting PeerWise as a summatively assessed component in introductory physics courses. Most report positive student engagement with and attitudes towards the system and several evidence a correlation between usage of the system and end-of-course outcomes (e.g., the final exam grade) [12,14,18,20,22] that impacts on students of all abilities in the course.

Far fewer studies have featured a discussion of the quality of questions authored by students. Hakulinen [13,14] and Korhonen [13] used expert ratings to investigate the reliability of the students' own quality ratings. They also explored automated approaches and expert sampling to classify the quality of questions on a binary basis ("good" vs "bad"), finding that the majority of student contributions were of good questions. Similarly, an investigation by Purchase *et al.* [23] of characteristics such as topic coverage, holistic question quality, difficulty, and indexing of questions submitted by a first-year programming class found that the question repository was of generally a high standard. Bottomley and Denny [16] report a study of a cohort of 107 second year biomedical science students. Over 90% of the contributed questions were found to be correct and of those incorrect questions, approximately half were identified as such by students who answered and/or commented on the question. The majority of questions contributed by students were classified at the lowest taxonomic level ("recall" or "remembering") with less than 10% above level 2 ("understanding") of Bloom's taxonomy. Bottomley and Denny state that this is to be expected, since for these students this was likely the first time they had been challenged to write their own questions. This finding is to be compared to similar studies that have used the same mapping procedure to categorize instructor-authored questions onto the levels of Bloom's taxonomy. One in particular presents surprising findings: a recent study [24] examining 9713 assessment items submitted by 50 instructors of introductory biology courses in the United States reported that 93% of the questions were at the lowest two levels of the revised Bloom's taxonomy. Zheng *et al.* [25] have applied the same procedure to provide evidence that the questions in MCAT examinations are strong from this perspective, but find similar high proportions of questions at the lowest levels of Bloom's taxonomy in other university examinations.

Given the widespread use of the PeerWise system, yet the paucity of reported studies as to the quality of contributed questions, it is both timely and necessary to address this issue. This paper reports a comprehensive evaluation of the question and explanation quality in student-authored questions across two separate, consecutive introductory physics courses delivered over two successive academic years at the University of Edinburgh. The paper is organized as follows: in the next section, we report brief details of the educational context of the courses and the specific details of PeerWise implementation in the courses, together with the *post-hoc* analysis procedure. We then present results of the question quality, mapping onto the levels of Bloom's taxonomy, and the quality of explanations using a classification rubric of our own devising. We briefly present initial findings in terms of analysis of numbers of student responses as a function of question quality before discussing our results more broadly. We end with some conclusions and suggestions for further research.

III. METHODOLOGY

The educational context for this intervention is a pair of consecutive introductory physics courses in the first year of the physics program at the University of Edinburgh. Physics 1A is a first course in classical mechanics and statics, covering kinematics, Newton's laws, energy, momentum, rotational motion, and oscillations. Physics 1B is a "showcase" course in modern physics, covering broad topics at an introductory level including quantum mechanics, thermal, nuclear, and particle physics. Both are taken by a broad and diverse student cohort in terms of prior study and future aspirations. Approximately half of the cohort are studying towards a physics degree. The remainder are taking the courses as an elective subject; these students are as equally qualified (in terms of high school physics and mathematics grades) as the physics majors. Approximately three-quarters of the cohort are male, and the vast majority of all students are aged between 17 and 19. More than 95% of the first semester Physics 1A cohort go on to take Physics 1B. Both courses have employed a variety of research-based and interactive engagement strategies, including extensive use of clickers, and studio-based workshop teaching [26] (similar to the TEAL [27] model). In Physics 1A, we have employed the FCI [28] at the start and end of the course to gauge both incoming cohort ability and effectiveness of instruction. Typical FCI results, which are relatively stable over several years, are a preinstruction cohort average of around 65% and a postinstruction normalized gain value of around 0.5–0.6. We have previously reported results [29] of pre-post CLASS [30] scores for the cohort, with the most significant finding being the high on-entry overall agreement with expert views (around 69%–72% over the past 4 years).

This study reports on data from the incorporation of PeerWise activities into the summative assessment of both the Physics 1A and 1B courses, over two successive academic sessions (2010–11 and 2011–12 academic years, hereafter referred to as 2010 and 2011 data for simplicity). The first of these two years was a pilot implementation: one weekly assessment task in each of Physics 1A and 1B was replaced by a PeerWise activity in which students were required to contribute, as a minimum, one original question that they authored, answer five others, and comment and rate on a further three questions. In Physics 1A 2010, the PeerWise activity was launched in week 5 of the semester, with the assessment deadline one week later. In Physics 1B 2010, an identical assessment requirement was set in week 4 of the semester, with a due date at the end of teaching in week 11. In each case, the PeerWise assessment contributed approximately 3% of the summative assessment grade for the course, with the scoring system built into the PeerWise system serving as the basis for allocation of assessment marks (for further details see Bates, Galloway, and McBride [22]). In the 2011 deployment in Physics 1A, we replaced three weekly assessments with PeerWise assessment activities (with one-week deadlines in weeks 5, 8, and 11 of the semester), and in Physics 1B 2011 just one weekly assessment activity, identical to the situation in 2010. Each of these individual tasks were the same as in the pilot: students were required to contribute, as a minimum, one authored question, to answer five others, and to rate and comment on a further three.

For both years of deployment, a significant component (90 min) of one of the class sessions in Physics 1A was devoted to preparatory activities ahead of the first PeerWise assessment activity of the year. (Only a handful of the Physics 1B students did not take Physics 1A in the previous semester, so for the vast majority of students these activities also informed their participation in the 1B PeerWise exercises.) These sessions, deliberately designed to help scaffold the process of writing questions of high quality, comprised four elements:

- A content-neutral quiz [31] that taught the language of MCQs (stem, options, key, distractors) and demonstrated how poorly written questions sometimes test nothing but language skills.
- A magazine-style self-diagnosis quiz to help students to explore their beliefs about thinking and guide them toward learning orientation and away from performance orientation.
- A question template introducing students to the notion of challenging themselves to write questions at a level just beyond their current understanding, aligned with the notion of operating in Vygotsky's "Zone of Proximal Development" [32]. This simplified constructivist model, along with information about common misconceptions and errors, encouraged the students to author questions of high cognitive value.
- A good quality example question, based on the template, which set a very high bar for expected creativity and complexity.

We devoted approximately 90 min of class time to covering these four elements, with a final activity in which students worked in groups of five or six to collectively author a question using our template. These group-authored questions were then uploaded to the online system to seed the database prior to setting the assessment task. All of our scaffolding materials are freely available online [33].

Analysis of question and explanation quality was conducted *post-hoc*. Two of the authors (J. R. and D. H.), undertaking final-year undergraduate honors projects and working in collaboration with the faculty member authors (S. P. B. and R. K. G.), devised and agreed on a series of three classifications to determine question quality. The first of these classified the cognitive level of the question based on the levels in Anderson and Krathwohl's revised version [4] of Bloom's taxonomy [3], illustrated and summarized in Table I. When undertaking this classification, the question itself, the question setter's provided

TABLE I. Categorization levels and explanations for the cognitive domain of Bloom's taxonomy.

Level	Identifier	Explanation and interpretation
1	Remember	Factual recall, knowledge, trivial "plugging in" of numbers.
2	Understand	Basic understanding, no calculation necessary.
3	Apply	Implement, calculate, or determine. Single topic calculation or exercise involving application of knowledge.
4	Analyze	Typically multistep problem; requires identification of problem-solving strategy before executing.
5	Evaluate	Compare and assess various option possibilities; often qualitative and conceptual questions.
6	Create	Synthesis of ideas and topics from multiple course topics to create significantly challenging problem.

solution, and subsequent comments in the question's comments thread were all visible to the rater: this wide range of information (and, in particular, the availability of the solution) made assigning a classification for cognitive level more straightforward than if the question alone had been available.

Bloom's taxonomy itself has received some criticisms, for example, by Moore [34]. These include reservations over the ordering of its hierarchy, and the challenges inherent in mapping between the cognitive level of assigned tasks and the actual processes of learning. Alternative frameworks could be used when classifying tasks, such as Illeris' three dimensions of learning [35], which distinguishes between external interaction processes and internal psychological processes, or an approach based on categorizing scientific practices such as analyzing data, constructing scientific arguments, or using mathematical models (for example, those practices identified by the NRC [36]). Nevertheless, we adopted a Bloom-based approach for a number of reasons: first, it facilitates a straightforward comparison with corresponding previous work in other disciplines [16]. Additionally, the precise nature of the hierarchy is not essential to our classification approach or findings; in fact, we found that the Bloom categories map clearly and naturally to distinguishable question types, but the actual ordering of the categories above level 3 is not critical. Furthermore, Bloom's taxonomy is widely known and understood, and in this case we use it exactly for its intended purpose (as Moore acknowledges [34]): classification of the objectives of tasks assigned to students.

It is important to distinguish clearly between this type of question classification and that conducted in some previous, well-known studies. For example, Chi, Feltovich, and Glaser [37] and Mason and Singh [38] have established that undergraduate students perform poorly in comparison to experts when they are required to classify physics problems by the physics principles required to solve them. In those studies, students were presented with physics problems (without solutions) and asked to categorize them by the underlying solution processes, i.e., they required students to identify how the problem could be successfully tackled. In our study presented here, the classifiers have access to both the problem and its solution, and do not have to identify a solution strategy; rather, they classify by the level of cognitive sophistication required by the suggested solution. This is a substantially more tractable process and one that can be performed much more reliably (as we establish later).

The second scale classified the quality of the explanation and solution associated with each question (which student contributors are required to provide at the time of authoring the question). This categorization scheme is illustrated in Table II.

Finally, overall criteria were devised to determine whether or not a question could be judged to be a high quality question. These criteria included required characteristics based on the cognitive level of the question (above factual recall, i.e., Bloom's level 2 or higher) and explanation quality ("minimal" or higher), together with other measures. A question was classified as high quality if it met all of the criteria outlined in Table III. Requiring at least two plausible distractors eliminates yes or no or true or false questions from being classified as high quality. In terms of

TABLE II. Categorization levels for explanation of solution to questions.

Level	Identifier	Description
0	Missing	No explanation provided or explanation incoherent.
1	Inadequate	Wrong reasoning and/or answer. Solution may be trivial, flippant, or unhelpful.
2	Minimal	Correct answer but with insufficient explanation or justification. Some aspects may be unclear or incorrect.
3	Good	Clear and sufficiently detailed exposition of both correct method and answer.
4	Excellent	Thorough description of relevant physics and solution strategy. Contains remarks on plausibility of answer and/or other distractors. Beyond normal expectations for a correct solution.

TABLE III. Criteria used to define high quality questions.

Measure	Criteria details
Taxonomy category	At least level 2 or higher (understand or above).
Explanation category	At least level 2 or higher (minimal or better).
Clearly worded question	Unambiguous vs unclear (binary measure).
Distractors	At least 2 feasible and plausible distractors.
Correctness	Most likely correct (binary measure).
Plagiarism	Not obviously plagiarized (binary measure).

identifying correctness and originality, the caveats not obviously (plagiarized) and most likely (correct) were added as it was not practical to cross reference all questions with all those publicly available (as end-of-chapter problems or on the web), nor to work through numerical solutions to all the problems contained within sampled questions. However, for a representative sample of questions, we pasted the question stem text into a search engine to check against material openly available on the internet and confirm the originality of the questions. We found indications of plagiarism from internet sources in only a negligible number of cases. Furthermore, the innovative contexts for many questions (e.g., see examples in the Supplemental Material [39]) are clearly different in style to those usually found as end-of-chapter problems.

For categorization of question cognitive level and explanation quality, interrater reliability tests were undertaken to ensure consistency between the coders. Each of the coders categorized a sample of questions from different course repositories for the cognitive level and explanation quality, before exchanging sample question sets and classifying each other's questions. The interrater reliability was subsequently determined by calculating Cohen's kappa [40]. Although there are several different methodologies to determine interrater reliability, there are a number of advantages to Cohen's kappa: it naturally accounts for expected purely coincidental agreement between raters, and handles discretized rating schemes well. It is usually considered a robust and rather conservative estimate of the interrater reliability. An initial sample of 35 questions produced moderately good interrater reliability. Following discussion between the raters and with the faculty member authors, a further 22 questions were sampled and coded from each repository. This yielded very good interrater reliability between all the contributors (over 90% for both question level and explanation quality).

In other projects, we have established that for undergraduate students working within our group on PER projects, the initial interrater reliability between student and expert (faculty) rater is over 70%. Following calibration and discussion of an initial set of ratings of questions, this level of agreement rises to over 90%. Furthermore, it is sustained at this level several weeks later without any further intervention or calibration in the mean time. This provides strong evidence that these students learn how to do this classification well and can undertake it very reliably.

TABLE IV. Number of questions sampled across all courses (percentages in brackets give the fraction of the course repository that was sampled in each case).

Course	2010	2011
Physics 1A	150 (42%)	200 (24%)
Physics 1B	179 (52%)	73 (46%)

Representative samples of the four distinct question repositories were coded, 602 questions in total. These were divided across the four courses as illustrated in Table IV. The absolute values give the actual number of questions coded from each repository, with the percentages illustrating what fraction of the total repository those numbers represent. As can be seen from this table, the 2011 repository for Physics 1B contains substantially fewer submissions than in 2010, despite the classes being of comparable size and the same task requirement being set. This corresponds to a reduced level of student submissions, which is possibly a negative consequence of increasing the load from 1 to 3 assignments in Physics 1A in 2011.

Contingency table analysis was used to identify differences in the distributions of the cognitive level or explanation quality of coded questions, between courses, or between years. The significances of any differences were calculated by performing χ-squared tests; we assigned statistical significance to test statistics with $p < 0.05$.

IV. RESULTS

In this section we present the results of classifying questions from the various course repositories. The Supplemental Material [39] for this paper presents case studies of a small number of student-authored questions, where we detail the rationale for why the questions and explanations were classified in the selected categories, and present brief details of the discussions that took place within the student cohort around these particular questions [39].

A. Cognitive level of questions

Categorizations of the cognitive level of questions sampled from repositories of one course (Physics 1A) over two successive years are shown in Fig. 1, based on the taxonomy level descriptions from Table I. For both years, there is only a small proportion of questions in the lowest taxonomy category (less than 5% for both years) and the majority of student questions are categorized as those requiring application or analysis, usually in the form of a quantitative problem to be solved in one or multiple stages, respectively. A χ-squared test indicates that there is

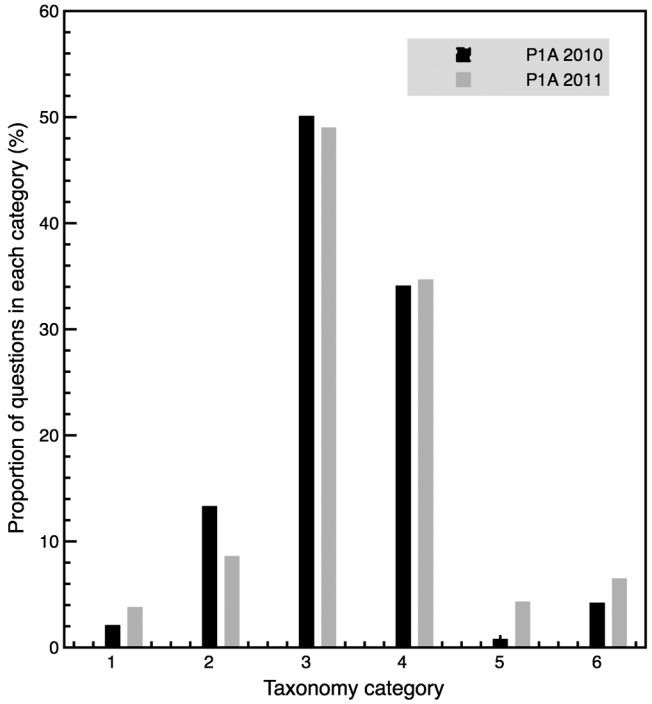

FIG. 1. Proportion of questions in each taxonomic category, for Physics 1A question repositories for 2010 ($N = 150$, dark bars) and 2011 ($N = 200$, light bars).

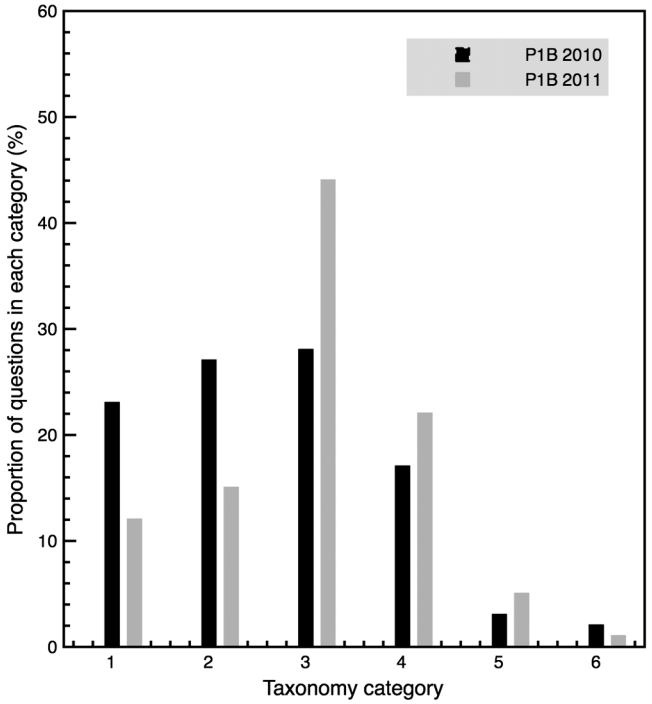

FIG. 2. Proportion of questions in each taxonomic category, for Physics 1B question repositories for 2010 ($N = 179$, dark bars) and 2011 ($N = 73$, light bars).

no statistically significant difference ($p = 0.27$) between the distributions from the two years.

Figure 2 presents the equivalent data for the set of questions sampled from the Physics 1B repositories over two successive years. Once again, we find a distribution spanning all taxonomic categories, but with some interesting differences from the distribution in Fig. 1. First, there is a difference in the shape of the profile between the two successive courses in both of the two years under study. Recall that the cohorts for Physics 1A and 1B in any given year are essentially the same group of students. This suggests that there is an influence played by the nature of course material on the types of questions students create. In the Physics 1A course (covering introductory topics in mechanics and oscillations) there is a different profile to the 1B course (designed as "grand tour" of modern physics, covering more material in somewhat lesser detail). This is not altogether surprising since it is natural to assume that the particular subject matter and design of the course will have some bearing on the nature of the student-authored questions.

In contrast to the results for Physics 1A, Fig. 2 exhibits a statistically significant difference ($p = 0.022$) between sampled 1B questions from subsequent years. In particular, there is a smaller fraction of questions in the lower taxonomic levels in the 2011 data. Since the 2010 and 2011 Physics 1B courses were very similar in most respects (and retained the same instructors), this may be a result of the different PeerWise implementation strategies in the two years. In 2010, a single PeerWise activity was introduced into Physics 1A. In 2011, three separate activities were undertaken in 1A. Thus, the 2011 1B cohort has had greater experience and practice in authoring questions. An interesting direction for future study would be to explicitly test this "quality improves with practice" hypothesis.

B. Explanation quality

A similar analysis was undertaken for the quality of student-authored explanations associated with each question, using the classification rubric shown in Table II. The PeerWise system does not *require* the explanation field to be completed prior to submitting the question into the repository. However, we made it clear to students that the ability to articulate the solution strategy, together with the ability to explain why incorrect answers were wrong, was a required and important part of developing a question.

Figure 3 shows data from questions sampled from the 2010 and 2011 Physics 1A question repositories. The figure shows that in over 95% of cases students did construct an explanation of some kind, and in the vast majority of cases these were of good or excellent quality (approximately two-thirds of the questions sampled were in the good or excellent categories for each of the two years). What is particularly impressive is the proportion of questions in the uppermost explanation category, those which went far beyond what might have been ordinarily expected

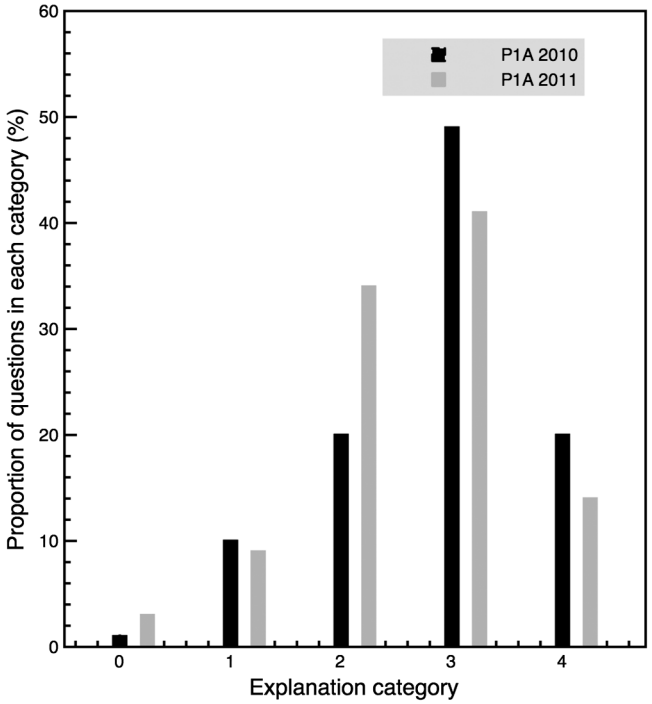

FIG. 3. Proportion of questions in each explanation category, for Physics 1A question repositories for 2010 ($N = 150$, dark bars) and 2011 ($N = 200$, light bars).

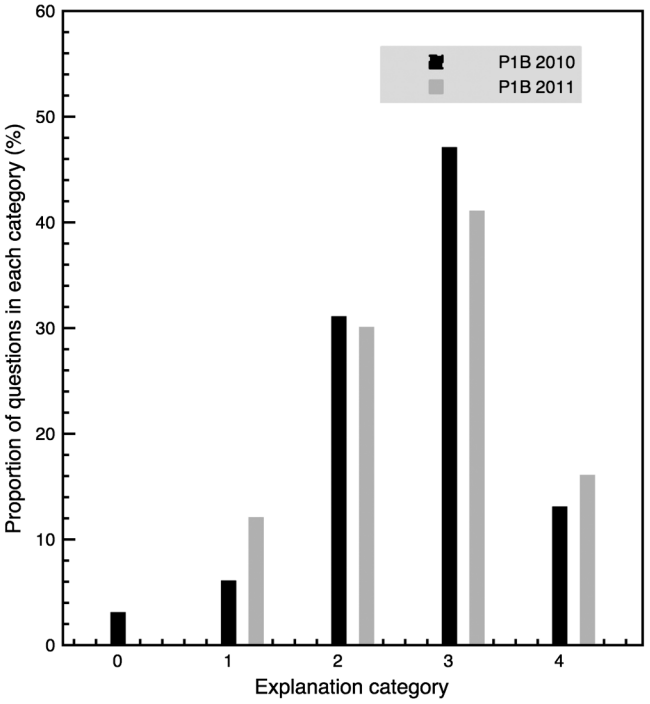

FIG. 4. Proportion of questions in each explanation category, for Physics 1B question repositories for 2010 ($N = 179$, dark bars) and 2011 ($N = 73$, light bars).

from a student solution to a problem. Many of the explanations in this category demonstrated important developmental skills within the discipline: sense making of the answer, appealing to special cases, multiple routes to solve the same problem, and articulation of commonly held alternate conceptions when discussing distractors.

A χ-squared test revealed a statistically significant difference between the distributions of the explanation categories for the Physics 1A data for 2010 and 2011 (shown in Fig. 3) with $p = 0.022$. The sampled explanations from the 2011 cohort appear to be of slightly lower quality overall compared to 2010 data. With hindsight, this pair of distributions confirms our experiences as instructors that the three separate PeerWise assessments within a single semester, when viewed "in the round" with all the other course assessment tasks, were probably too many, evidenced by a noticeable drop in the proportion of students engaging with the third assessment in Physics 1A in 2011 (data not presented here) and also in Physics 1B in 2011 as previously noted. Notwithstanding, the overall profile of both distributions is striking, with very few questions lacking meaningful explanation. The modal classification was in the good category (between 40% and 50% of all questions in each course sample) and a non-negligible fraction were in the highest category in both cases.

Equivalent data for the explanations associated with questions sampled from the 2010 and 2011 Physics 1B repositories are shown in Fig. 4. Here, there is no statistically significant difference ($p = 0.66$) between samples for two successive years, and the same overall pattern in the profile of explanation quality detailed above for the 1A questions is once again clearly apparent.

C. Overall question quality

To be classified overall as a high quality question, a submission was required to meet *all* of the criteria presented in Table III. These include minimum requirements for the classification of cognitive level and explanation quality (at least level 2 or higher in both cases: *understanding* or above in terms of cognitive level, and a *minimal* level of explanation or above). In addition, these criteria also included further quality requirements pertaining to question clarity, plausibility of distractors, originality, and correctness.

Combining all sampled questions together ($N = 602$), our classification yielded that overall 453 questions (75%) met all the criteria outlined in Table III. In terms of individual criteria failure rates, our analysis showed the following:

- On grounds of clarity, only 5% of questions were rejected on the basis that the question statement was unclear, ambiguous or irrelevant;
- In terms of the number of distractors, 80% of questions had at least two plausible distractors;
- 10% of questions were rejected on the basis of having insufficient explanation;
- 10% of questions were rejected on the basis of too low a taxonomy classification;

- Only 1% of questions were identified as being obviously plagiarized (and in most cases, these were highly derivative of questions already in the repository, or of problems elsewhere in the course materials);
- Only 5% of questions were rejected because they were identified as incorrect or seriously flawed in some way (and in more than half of these instances, the error or mistake had been identified by other students).

D. Student answer patterns as a function of question quality

An extensive treatment of student behavior in terms of question answering and the related comments or discussions is beyond the scope of this paper. However, we do present data to address the question: "Do students answer easier questions more frequently, and if so by how much?" One might imagine a situation where students fulfill minimum assessment requirements by strategically targeting "easy" questions, thus putting in the minimum effort.

For the 602 questions evaluated across the four course offerings, we have calculated the mean number of answers per question as a function of question taxonomic category. Representative data are shown for Physics 1A 2010 and 2011 in Fig. 5. With a relatively small number of distinct course data sets (4), we must, of course, be cautious about drawing too many conclusions. Nonetheless, the data suggest that questions in lower taxonomic categories tend to be answered more, but only by a relatively small factor. For example, in questions sampled from the Physics 1A 2011 repository, category 1 and 2 questions are answered on average 21(4) and 18(4) times, respectively, where the values in parentheses give the standard error on the mean. For questions in higher categories, the mean number of answers per question ranges from 8(1) to 11(7), though for smaller numbers of questions in higher categories the statistics can be somewhat distorted by one or two questions that tend to be extremely popular. The two different courses (Physics 1A and 1B) show broadly similar behavior: that is, questions in higher taxonomic categories tend to attract fewer answers per question. In summary, as illustrated by Fig. 5, there is evidence for reasonable consistency in patterns of answering in the same course over successive years: students answer the lower category questions more frequently, but do also answer a substantial number of the longer, more involved and more challenging higher category questions.

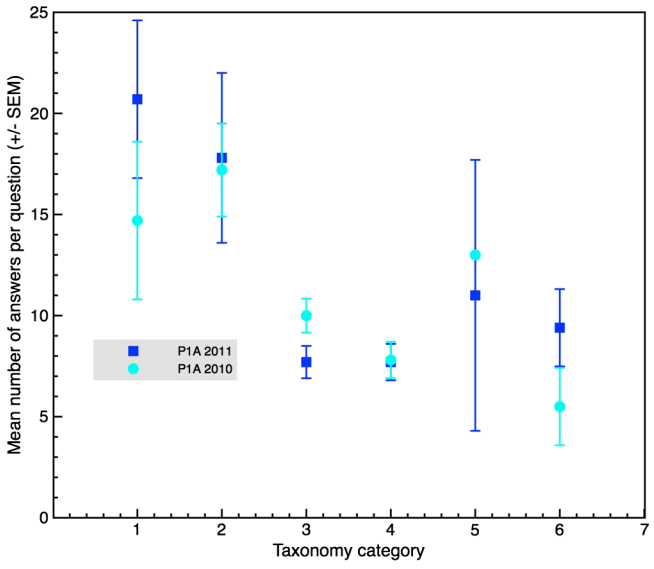

FIG. 5. Mean number of student responses per question as a function of question taxonomy, for Physics 1A course repositories for the 2010 and 2011 cohorts. Error bars denote the standard error on the mean. Note that for P1A 2010, there was only one category 5 question in the sample.

V. DISCUSSION

The quality of the student-authored questions examined in this study, in terms of their categorization onto corresponding skills in the cognitive domain of Bloom's taxonomy, is significantly different than that reported previously in different subject domains. Bottomley and Denny [16] report that 90% of question items authored by biology undergraduate students lie in the lowest two taxonomy levels, with more than half in the lowest category. In contrast, we report a broad distribution across all taxonomic levels, with a majority of the questions in the middle categories of application and analysis. As some critics of Bloom's taxonomy observe, categorization can be a somewhat subjective activity: we have attempted to ensure that our process is as robust as possible with appropriate interrater reliability checks. However, it is certainly still possible that what one person might interpret as being appropriately categorized as analysis may be better classified as evaluation by someone else. That said, it is rather more straightforward, with a knowledge of the course material covered, to determine those questions that really are in the lowest category of factual recall rather than belonging higher up the classification. This is one reason why we have adopted the minimum criterion of taxonomy classification to be understand or higher for the question to be potentially judged as a high quality one: we may debate a question categorization at the highest levels, but in our experience everyone can consistently recognize one at the lowest.

There are also similarities to be drawn with the study reported by Bottomley and Denny [16]. They too find that over 90% of questions are accompanied by an adequate explanation or better (rated on a four point scale, rather than our five point one). Likewise, they find similarly high proportions of correct solutions and not-obviously-plagiarized material as we report here. There are a number of possible explanations of why we find a higher quality of student-authored questions in the present study. First, the

research-based curriculum found in both Physics 1A and 1B features a large number of more cognitively advanced questions, both as worked examples and as student tasks. These should provide a model for relatively sophisticated types of physics problems. Nevertheless, question setting will be an unfamiliar task for most of the students; it is our hypothesis that the higher quality of student-authored questions found in the present study may be connected to the introductory exercises and scaffolding activities that we provided to students ahead of the first PeerWise assessment task. These not only serve to set the bar at a high level in terms of expected contributions (by provision of a high quality worked example), but also provide support for (many, but by no means all) students to extend themselves beyond what they currently know, thus challenging their own understanding. Ideally, to test this hypothesis we would seek to set up a controlled experiment that contrasts question quality from a control and two intervention groups (no PeerWise activity, PeerWise with no scaffolding, and PeerWise with scaffolding). We have not yet gathered such data, but do note that previously reported studies of question quality do not describe significant scaffolding or introductory activities. (Previous implementations of PeerWise which do report the use of scaffolding [19,20] found enhanced levels of student engagement but did not investigate the quality of contributed questions.) Analysis of replication studies in other disciplines and institutions, utilizing similar scaffolding materials modified appropriately for local contexts, is under way and will be reported elsewhere.

Given that we find a distinctive pattern of high level question setting by students when analyzed using a classification based on Bloom's taxonomy, it might also be instructive to investigate alternative categorization approaches which examine different characteristics to those of Bloom (and which might not share some of its limitations). One potential approach would be to classify student-generated questions based on which scientific practices they motivate: this would distinguish between, for example, questions which merely require recall and those which necessitate model building or analysis of data, etc.

Our results indicate that not only do students produce, on the whole, very good questions, but also the appropriately detailed and useful explanations to accompany them. It may be the case that having ownership of a question encourages students to create more detailed commentary, as does the responsibility of contributing to a resource that will benefit their peers, plus the positive feedback received through comments and "badges" within the system. Informal consultations with students suggested that the median time to create and refine a question plus develop a solution was between 1 and 2 hours. Though there is considerable variation in how much time students spend engaged with the system, this seems an appropriate time investment for the summative assessment credit (typically 2%–3% of course grade) associated with each PeerWise task.

We find differences in the distribution of questions across the taxonomic levels for different courses, but far less difference between different years of the same course (even though in the case of Physics 1A the use of PeerWise was increased from one to three activities in 2011). This is perhaps not surprising as the particular course content material will have a bearing on the types of questions students author. Physics 1A is a first course in classical mechanics, with rich contextualization possibilities from everyday scenarios. Physics 1B is a grand tour course of the fundamentals of modern physics, with a greater range of topics covered (thus in somewhat less depth); there are fewer obvious avenues for real-world contextualization of questions in this course as compared to Physics 1A. Despite these differences, the same broad conclusion of students being very capable of producing high quality questions (and explanations) holds.

Our results suggest that questions that are classified in the lowest two taxonomic categories tend to be answered more frequently, despite there generally being fewer of them. One possible explanation for this observation is the time required to answer each type of question: a category 1 or 2 question could be solved by a few seconds' careful thought, enabling students to make rapid progress through them, whereas higher category questions might require tens of minutes of calculation and problem solving. The difference in mean number of answers per question as a function of taxonomic category is around a factor of 2, which is somewhat smaller than might be expected if a large fraction of the cohort are being tactical and trying to answer the most straightforward questions for "easy marks." We emphasize the significance of this observation: the students *could* have satisfied the stipulated task by rapidly answering predominantly low-level, straightforward questions with minimal investment of time and effort. The fact that they also answer more sophisticated questions—requiring orders of magnitude more time and higher cognitive demands—still at around half the frequency of the lower-level questions illustrates that the students are engaging substantially with answering the more sophisticated questions as well as writing them. This factor is also likely to be heavily influenced by the particular requirements for an assessed task: insisting that students answer a large number of questions as a minimum requirement is likely to lead to more tactical choice of questions to answer. Likewise, requiring students to set an unreasonable number of questions in proportion to the time or effort they have available is highly likely to result in questions of lower quality. Thus, we would suggest that the context in which the system is used in a course, together with the material and support provided to help students write questions, are both important factors that have a bearing on question quality.

VI. CONCLUSIONS AND FUTURE WORK

We have classified student-authored questions produced as part of the summative assessment for four introductory physics courses (two semester-long courses, over two successive academic sessions) according to cognitive level and quality of explanation. We find that these first-year students are capable of producing very high quality questions and explanations. On the basis of minimum thresholds for cognitive level and quality of explanation, together with other question-specific criteria, we find that 75% of the questions can be classified as being of high quality. Questions meeting these criteria are clear, correct, require more than simple factual recall to answer, and possess a correct solution and plausible distractors. In particular, a substantial fraction of the questions constitute true problems (as opposed to simple exercises). A significant difference between our implementation of PeerWise and other reported studies examining contributed question quality is the provision of support and scaffolding materials prior to the start of the assessment activity.

Our previous work with PeerWise has demonstrated that incorporation into the summative assessment strategy for a course can lead to good engagement and enhanced learning [22], confirmed by similar studies in other disciplines [12,14,18,20]. The present work complements these studies, indicating that students are capable of producing high quality questions and detailed explanations. These findings, coupled with the efficiency associated with student assessment being largely done by their peers, suggest that this instructional methodology can become part of the "standard toolkit" of student-centered course design for undergraduate physics.

This remains a fruitful area for on-going research, including replication studies at different institutions and in different subject contexts, which we will report elsewhere. Further detailed analysis is under way of the student comments, and more broadly a learning analytics or student network analysis to understand interactions between question authors and answerers. Finally, given that the proportion of students taking introductory physics courses far exceeds those going on to study for a physics degree, we are investigating the incorporation of such strategies into classes composed entirely of nonmajors. Our work with PeerWise in undergraduate classes suggests that students are substantially more creative than we might have previously given them credit for, and this creativity might be usefully harnessed in meaningfully developing core skills (such as problem solving) within the discipline.

ACKNOWLEDGMENTS

This work has been partly funded under a grant from the Joint Information Systems Committee (JISC), under their Assessment and Feedback strand. Significant contributions to the project have been made by other members of the Physics Education Research Group at the University of Edinburgh, including Judy Hardy, Karon McBride, and Alison Kay.

[1] S. Draper, Catalytic assessment: Understanding how MCQs and EVS can foster deep learning, Br. J. Educ. Technol. **40**, 285 (2009).

[2] M. Bieber, J. Shen, D. Wu, and S. R. Hiltz, Participatory learning approach, in *Encyclopedia of Distance Learning*, Vol. 3 (IGI Global, Hershey, PA, 2005), pp. 1467–1472.

[3] B. S. Bloom, *Taxonomy of educational objectives: The classification of educational goals; Handbook I: Cognitive Domain* (Longman, New York, 1956).

[4] L. W. Anderson and D. R. Krathwohl, *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives* (Longman, New York, 2001).

[5] M. Smith and K. Perkins, At the end of my course, students should be able to..: The benefits of creating and using effective learning goals, *Microbiol. Australia* **31**, 35 (2010).

[6] The 1000 most visited sites on the web, (16 Feb 2013), http://www.google.com/adplanner/static/top1000/.

[7] Wikipedians, (16 Feb 2013), http://en.wikipedia.org/wiki/Wikipedia:Wikipedians.

[8] P. W. Foos, Effects of student-written questions on student test performance, Teach. Psychol. **16**, 77 (1989).

[9] D. J. Nicol and D. Macfarlane-Dick, Formative assessment and self-regulated learning: A model and seven principles of good feedback practice, Studies High. Educ. **31**, 199 (2006).

[10] D. J. Nicol, E-assessment by design: Using multiple-choice tests to good effect, J. Further and Higher Educ. **31**, 53 (2007).

[11] P. Denny, A. Luxton-Reilly, and J. Hamer, The PeerWise system of student contributed assessment questions, in *Proceedings of the tenth conference on Australasian computing education - Volume 78, ACE '08* (Australian Computer Society, Inc., Darlinghurst, Australia, 2008), pp. 69–74.

[12] P. Denny, A. Luxton-Reilly, J. Hamer, and H. C. Purchase, PeerWise: students sharing their multiple choice questions, in *Proceeding of the Fourth International Workshop on Computing Education Research* (ACM, New York, 2008), pp. 51–58.

[13] L. Hakulinen and A. Korhonen, Making the most of using PeerWise in education, in *ReflekTori 2010 - Symposium of Engineering Education, 2010. Aalto University, Lifelong Learning Institute Dipoli* (Aalto University, Espoo, Finland, 2010), pp. 57–67.

[14] L. Hakulinen, Using Computer Supported Cooperative Work Systems in ComputerScience Education - Case: PeerWise at TKK, Master's thesis, Faculty of Information and Natural Sciences, School of Science and Technology, Aalto University, 2010.

[15] P. Denny, B. Hanks, B. Simon, and S. Bagley, *PeerWise: Exploring conflicting efficacy studies,in Proceedings of the seventh international workshop on Computing education research, ICER '11* (ACM, New York, 2011), pp. 53–60.

[16] S. Bottomley and P. Denny, A participatory learning approach to biochemistry using student authored and evaluated multiple-choice questions, Biochem. Mol. Biol. Educ. **39**, 352 (2011).

[17] T. Hooper, S. Park, and G. Gerondis, Student perceptions of PeerWise Web 2.0 technology, in *TERNZ 2011 Conference, Victoria University of Wellington* (2011).

[18] A. Luxton-Reilly, D. Bertinshaw, P. Denny, B. Plimmer, and R. Sheehan, The impact of question generation activities on performance, in *SIGCSE 12: Proceedings of the 43rd ACM Technical Symposium on Computer Science Education, ACM SIGCSE* (ACM, New York, 2011), pp. 391–396.

[19] J. H. Paterson, J. Devon, J. McCrae, D. C. Moffat, and E. Gray, Enhancing the quality of student-generated MCQ's: A final report, 2011.

[20] A. Sykes, P. Denny, and L. Nicolson, *PeerWise - the Marmite of veterinary student learning*, in *Proceedings of the 10th European Conference on E-Learning, Vols 1 and 2*, edited by S Greener and A Rospigliosi (Academic Conferences Ltd., Curtis Farm, Kidmore End, NR Reading, RG4 9AY, England, 2011), pp. 820–830; *10th European Conference on e-Learning (ECEL), Univ Brighton, Brighton Business Sch, Brighton, England, 2011*.

[21] J. Paterson, J. Wilson, and P. Leimich, Uses of peer assessment in database teaching and learning, in *Data Security and Security Data*, Lecture Notes in Computer Science, Vol. 6121, edited by L. MacKinnon (Springer, Berlin, Heidelberg, 2012), pp. 135–146.

[22] S. P. Bates, R. K. Galloway, and K. L. McBride, Student-generated content: Using PeerWise to enhance engagement and outcomes in introductory physics courses, in 2011 Physics Education Research Conference, AIP Conf. Proc. **1413**, 123 (2012).

[23] H. Purchase, J. Hamer, P. Denny, and A. Luxton-Reilly, *The quality of a PeerWise MCQ repository*, in *Proceedings of the Twelfth Australasian Conference on Computing Education - Volume 103, ACE '10* (Australian Computer Society, Inc., Darlinghurst, Australia, 2010), pp. 137–146.

[24] J. L. Momsen, T. M. Long, S. A. Wyse, and D. Ebert-May, Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills, CBE Life Sci. Educ. **9**, 435 (2010).

[25] A. Y. Zheng, J. K Lawhorn, T. Lumley, and S. Freeman, Application of Bloom's taxonomy debunks the"MCAT myth", Science **319**, 414 (2008).

[26] S. P. Bates, Reshaping large-class undergraduate science courses: The weekly workshop, *International Journal of Innovation in Science and Mathematics Education* **14**, 1 (2005).

[27] Y. J. Dori, J. Belcher, M. Bessette, M. Danziger, A. McKinney, and E. Hult, Technology for active learning, Mater. Today **6**, 44 (2003).

[28] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).

[29] S. P. Bates, R. K. Galloway, C. Loptson, and K. A. Slaughter, How attitudes and beliefs about physics change from high school to faculty, Phys. Rev. ST Phys. Educ. Res. **7**, 020114 (2011).

[30] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).

[31] Designing good multiple choice questions, (16 Feb 2013), the quiz was adapted from the 'Grunge Prowkers' quiz by Phil Race, which is freely available online. http://www.slideserve.com/justus/designing-multiple-choice-questions-and-feedback-responses.

[32] S. Chaiklin, The zone of proximal development in Vygotsky's analysis of learning and instruction, in *Vygotsky's Educational Theory and Practice in Cultural Context* (Cambridge University Press, Cambridge, England, 2003).

[33] PeerWise scaffolding resources used in Physics 1A and 1B classes at the University of Edinburgh, (16 Feb 2013), http://www.peerwise-community.org/resources/#activities.

[34] D. S. Moore, Reconsidering Bloom's taxonomy of educational objectives, cognitive domain, Educ. Theory **32**, 29 (1982).

[35] K. Illeris, *The Three Dimensions of Learning: Contemporary Learning Theory in the Tension Field Between the Cognitive, the Emotional and the Social* (NIACE, Malabar, Florida, 2002).

[36] National Research Council of the National Academies, *A Framework for K-12Science Education: Practices, Crosscutting Concepts, and Core Ideas* (Natl. Acad. Press, Washington, DC, 2012).

[37] M. T. H. Chi, P. J. Feltovich, and R. Glaser, Categorization and representation of physics problems by experts and novices, Cogn. Sci. **5**, 121 (1981).

[38] A. Mason and C. Singh, Assessing expertise in introductory physics using categorization task, Phys. Rev. ST Phys. Educ. Res. **7**, 020110 (2011).

[39] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevSTPER.10.020105, which contains selected question examples, student discussion comments and rationale for categorization.

[40] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. **20**, 37 (1960).