J Sci Educ Technol (2016) 25:561–574
DOI 10.1007/s10956-016-9613-9

The Psychometric Properties of Classroom Response System Data: A Case Study

Gerd Kortemeyer[1]

Published online: 8 March 2016

Abstract Classroom response systems (often referred to as "clickers") have slowly gained adoption over the recent decade; however, critics frequently doubt their pedagogical value starting with the validity of the gathered responses: There is concern that students simply "click" random answers. This case study looks at different measures of response reliability, starting from a global look at correlations between formative clicker responses and summative examination performance to how clicker questions are used in context. It was found that clicker performance is a moderate indicator of course performance as a whole, and that while the psychometric properties of clicker items are more erratic than those of examination data, they still have acceptable internal consistency and include items with high discrimination. It was also found that clicker responses and item properties do provide highly meaningful feedback within a lecture context, i.e., when their position and function within lecture sessions are taken into consideration. Within this framework, conceptual questions provide measurably more meaningful feedback than items that require calculations.

Keywords CRS · Clickers · Classical Test Theory · IRT · Item Response Theory · Psychometrics · Physics

✉ Gerd Kortemeyer
kortemey@msu.edu

[1] Department of Physics and Astronomy, Lyman Briggs College, Michigan State University, East Lansing, MI 48824, USA

Introduction

The effectiveness of Peer Instruction (Mazur 1997) has been the subject of a number of studies (e.g., Refs. Crouch and Mazur 2001; Fagen et al. 2002; Lasry et al. 2008; Barth-Cohen et al. 2015), to name but a few, and while specific outcomes may be implementation dependent (Keller et al. 2007; Turpen and Finkelstein 2009; Beatty and Gerace 2009; Richardson et al. 2014), it is virtually undisputed among education researchers that this activating strategy is superior to purely transmission-style lectures. While in principle, Peer Instruction can be implemented using low-tech means such as flash cards (Lasry 2008), classroom response devices, colloquially referred to as "clickers," are an enabling technology for scalable and efficient deployment even in large-enrollment courses.

Like other research-based teaching innovations, clickers are and should be disruptive to the flow of a lecture: Instruction gets "interrupted" by periods of discussions among students, and based on the outcome of questions, instructors may be "forced" to change the emphasis or even the topic of the day "on-the-fly." This element of insecurity introduced by relinquishing control of the classroom instruction may be one of the reasons for the discomfort experienced by both students and instructors, which combined with the extra planning and writing effort (Caldwell 2007; Kay and LeSage 2009) may be one of the reasons for the slow adoption or even abandonment of Peer Instruction and other activating strategies (Dancy and Henderson 2010; Henderson et al. 2012). Particularly faculty members who never tried using clickers frequently voice concerns over the validity of the responses (Lantz 2010), assuming that the "noise" created by random answers would overshadow any possible insights. At the root of this argument may be a fundamental misunderstanding of the role of clickers during instruction: In the framework of Peer Instruction and other activating methods, clickers are not a testing tool, instead they are a teaching tool.

The validity of test questions, which are summative in nature, is generally analyzed using psychometric techniques, which assess the interaction of test questions with groups of examinees to arrive at quality measures such as difficulty and discrimination (Nunnally and Bernstein 1994). Clicker questions during instruction, which are formative in nature, do not need to have the same psychometric properties as examination questions, since they serve a different purpose.

Nevertheless, given the wide range of implications of using classroom response systems, it is appropriate to investigate how meaningful these responses are: Do they indeed reflect the understanding of concepts? How meaningful is the formative feedback that both instructors and students receive from clicker questions? Important measures are:

- Correlation to examination performance: Examinations should reflect the learning goals of a course, and an important question is how well clicker performance predicts examination performance. For both instructors and students, this is an important component of the predictive validity of clicker items (the word "item" in this context denotes what physicists would usually call a "problem").
- Reliability: Scores on clicker questions should have some level of internal consistency, e.g., it should be expected that high-ability students perform well across a number of questions. Reliability is a global measure that is related to the individual item discrimination.
- Discrimination: In line with their formative nature, clicker questions should both test and develop understanding of course concepts, which is only effective if they distinguish between students who understand the concept and those who do not.

These measures will depend on the particular implementation details of clicker usage, and no general statements can be made. We can, however, carry out a case study in a typical introductory physics course and begin to answer a number of questions: How much of the feedback is tainted by the typical low-stakes setting of Peer Instruction, which can lead to random guessing? And how much are these data systemically "tainted" by the Peer Instruction process, which in an examination setting would amount to copying or "cheating?" Do students and instructors get a false sense of security? While extensive research exists on clicker usage, these questions still remain largely open.

The data stream generated by clickers has been previously investigated with respect to gender differences in participation and effectiveness (King and Joshi 2008), as well as response timescales and modification of answer choices during polling (Richardson and O'Shea 2013). It was found that male students tend to participate less, but if they participate, they gain more in terms of examinations grades; this is somewhat surprising, since their answer choices appear to be more haphazard than those of female students, i.e., male students change their mind more frequently while polling is open. It is unclear whether this behavior taints the data gathered during lecture: Are students' initial or final choices more meaningful when it comes to assessing understanding of the subject matter, or is the mere fact that the students changed their minds in the first place indicative of vague conceptualization?

When it comes to clicker responses reflecting student learning and ability, an important difference appears to exist between anonymous and assigned clicker usage. If the instructor can identify which student submitted which answer, the percentage of correct responses increases (Poole 2012). Another implementation detail is whether or not to assign points for participation, only for correct answers, or for a combination of both (White et al. 2011). While the quality of individual answers may increase in such partial credit scenarios, there is evidence that the quality of peer discussions might suffer, as strong students tend to dominate in grade-conscious discourse (James 2006). In any case, the contribution of clicker performance to the total grade in courses is usually low, in the range of just a few percentage points. As opposed to examinations and even homework, which traditionally much more strongly influence the course grade, students in such low-stakes scenarios may not be on their best behavior: They may be guessing or choosing to not even read the question and think about the answer. If this happens, psychometric measures suffer (Setzer et al. 2013). Thus, performance may not necessarily be a true reflection of the learner's understanding of the topic and ability.

When investigating clicker question validity, deep insight can be gained from interviews (Ding et al. 2009) or listening in to student conversations (James and Willoughby 2011). Unfortunately, these approaches are time-consuming and not scalable when attempting to provide clicker questions for every lecture session during a semester. In this case study, standard psychometric techniques are applied to "clicker" data. After introducing the course setting (Sect. 2), the study narrows its focus from a global view of several semesters (Sect. 3.1) to a more in-depth look at one semester (Sects. 3.2, 3.3), and eventually to one lecture session (Sect. 4).

Course Setting

The study was carried out in introductory calculus-based physics courses, which were mostly taken by life science and pre-medical students. Data were available from four courses; three of the courses were first-semester mechanics with 200–250 students (split into smaller sections) and 70–250 clicker items, and one course was second-semester electricity and magnetism with 107 students and 200 clicker items. Sections met three times a week for an average of 18 weeks with interruptions by holidays and examinations, resulting in an average of about five clicker questions per lecture session—however, variations on this average are wide, particularly when questions are repeated after Peer Instruction. These courses were taught by two different instructors; however, both instructors followed the same classroom pedagogy. Clicker items were written by the course instructors simply to support the lecture content. While some collections of "good" clicker questions were used to find inspiration and reduce the load of coming up with new items (e.g., Ref. Mazur 1997), generally a less scientific and more pragmatic approach was used to just write questions as needed. Examinations in the course were multiple choice, with a mixture of conceptual and calculation problems, and determined the majority of the course grade.

In-class clicker performance contributed 5 % to the final course grade in all semesters; however, there were varying schemes for rewarding correct answers: In earlier semesters, incorrect answers received $\%I = 60\,\%$ credit, while correct answers received $\%C = 100\,\%$, but the students only needed 40 % of the total available points to receive 100 % credit. The reason for this low bar on getting full credit was a hesitancy on the part of the instructor to demand participation in this, at the time, "new" mode of instruction. In later semesters, an easier scheme was implemented, where incorrect answers received $\%I = 60\,\%$ credit and correct answers $\%C = 140\,\%$. In all semesters, each lecture session was evaluated separately, so that each lecture session had equal weight in the end, regardless of number of questions asked. In particular, if N questions were asked during a lecture session, and a student answered c correctly and i incorrectly, the credit for the session would be $(c \cdot \%C + i \cdot \%I)/N$. The grading model thus was a mixture between participation and correctness rewards, but was clearly low stakes.

Students had to purchase their own iClickers (2003) and personally register them in the LON-CAPA (Kortemeyer et al. 2008) course management system. Lecture attendance was very high throughout the semester, with typically only a handful of students being absent, and generally all attending students were also answering all of the clicker items. This is typical for pre-medical students when grade incentives are given, but may not be typical for other physics courses. Only multiple choice questions with answers from A up to E were posed, even though the iClicker system would have allowed for more complex question types. Students typically answered items individually when they were posed for the first time, but some students would talk to their neighbors while the poll was open. As the iClicker system for each poll records both the initial and final answer, it became apparent that students frequently changed their minds; the dynamics of this behavior are complex (see, e.g., Richardson and O'Shea 2013), and this study will only empirically analyze the properties of these initial and final answers. Depending on the final answer distribution, the instructors may or may not have asked students to discuss the question with their neighbors and then take a second or even third poll, following the Peer Instruction pedagogy (Mazur 1997).

Course Level Measures

In the following subsections, increasingly more factors are taken into account to gain measures of clicker feedback meaningfulness. Sect. 3.1 investigates the global correlation between clicker and examination scores for four courses. Section 3.2 starts taking into account properties of the question items by using Classical Test Theory within one of the courses, and Sect. 3.3 adds traits of the learners by employing Item Response Theory.

Correlation Between Examination and Clicker Correctness

A first approach to investigating how well clicker responses reflect student learning is to compare performance on clicker questions with performance on examination questions. To that end, all clicker and examination responses were collected over the course of the semester (i.e., from all lecture sessions, midterms and the final examination), and the fractional correctness on clicker and examination questions correlated. For clicker questions, both the initial response (first click) and the final response (last click while collection is open) were investigated to see whether either of these responses is more meaningful (if a student only clicked once while the collection was open, both responses are identical). Figure 1 shows the result.

For all four courses, the coefficients of determination R^2 fall into the range of 0.2–0.3, indicating a modest correlation. Interestingly, the final response is no better indicator than the initial response, instead it is simply more likely to be correct for all students, regardless of overall

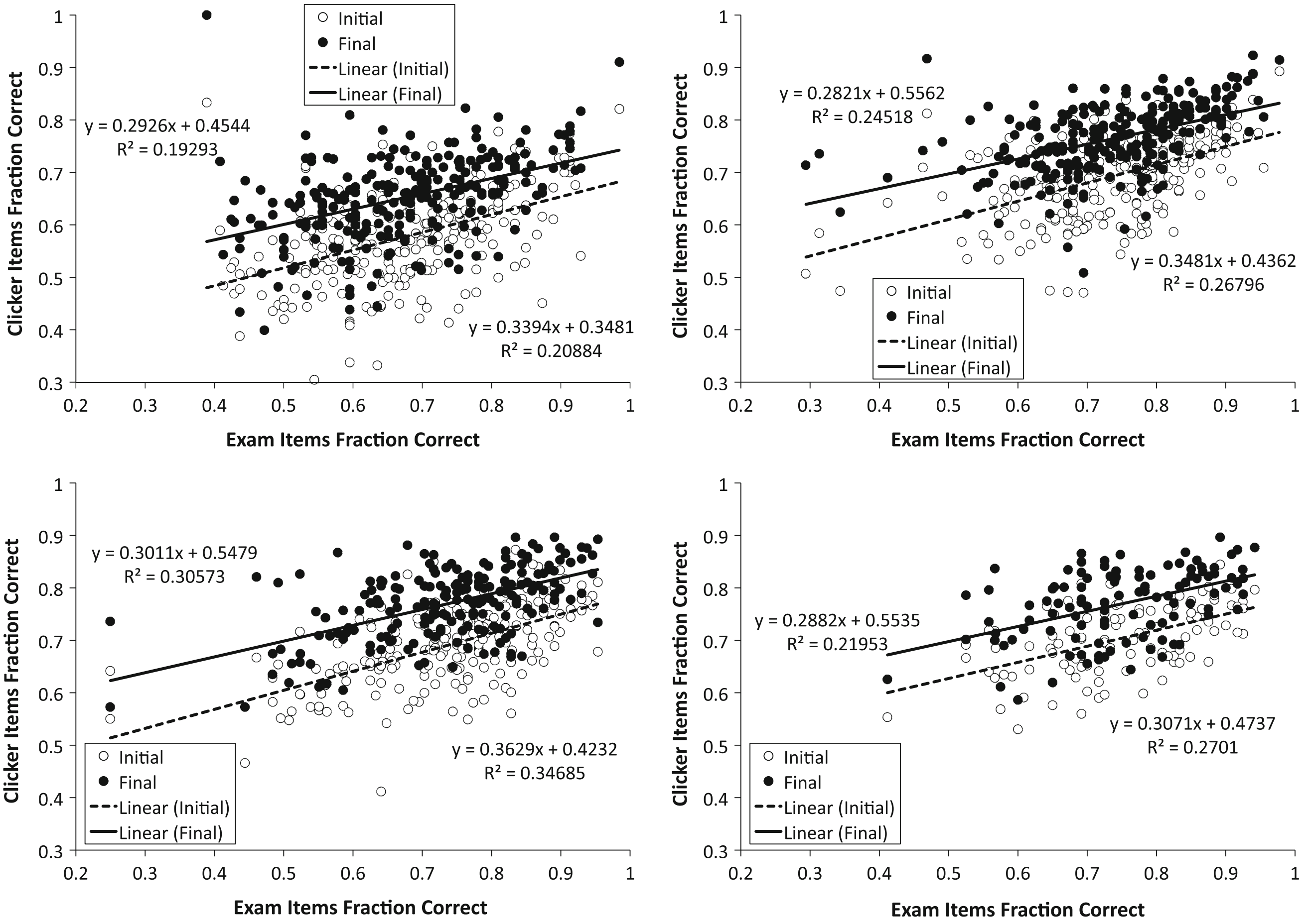


Fig. 1 Correlation between fraction of correct problems on examinations and fraction of correct answers to in-class clicker questions for four different introductory physics courses. Each data point represents one student, where *open circles* denote initial answers and *solid circles* denote final answers. Linear regression lines, as well as associated equations and R^2 values, are given for initial answers (*dashed*) and final answers (*solid*)

performance. Overall, clicker data have moderate predictive validity with respect to summative assessment outcomes.

To better understand this result, it is important to assess the quality of this "test" and its items. Tests have to be internally consistent. Good formative assessment problems have medium difficulty: They are not too hard, so they do not frustrate the majority of learners, but they are also not so easy to be meaningless. They also have high positive discrimination, so they give meaningful feedback to both learners and instructors. An item with negative discrimination is generally unusable: Low-ability students have a better chance of solving it than high-ability students (maybe due to a subtle difficulty that lower-ability students overlook, some component that makes high-ability students overthink the problem, or simply due to an error); "trick questions" can also have negative discrimination.

Unfortunately, based on our data set, we were generally unable to match performance on specific clicker items with specific examination items on the same topic, since for the vast majority of available data, we had the scores, but not the associated questions at our disposal. However, even if these data were available, the analysis would have been cumbersome at best and arbitrary at worst: In physics, concepts are very closely connected and build up over time. Clicker items often focus on one particular concept or even one facet of a concept, while examination items typically require the application and translation of multiple concepts. For example, a second-semester examination item on magnetism may include first-semester concepts of angular motion and energy conservation. We will, however, slowly "zoom in" on particular questions over the course of our analysis.

The following subsections will focus on one of the courses (249 students, 143 items) and deploy both Classical Test Theory and Item Response Theory to assess the psychometric properties of clicker data.

Classical Test Theory

Classical Test Theory (CTT) evaluates item characteristics such as difficulty and discrimination, as well as the reliability of tests, where in this case, the "test" is the outcome of the clicker assessments. Calculations are performed using the Classical Test Theory package (Willse 2014) within the R statistical software system (2008).

As opposed to clicker deployment, examination settings are highly controlled, and instructors might spend more time designing and proofreading examinations than clicker items. Thus, a first question is whether or not this is reflected in a comparison between these two kinds of "tests." The left panels of Fig. 2 show the *P* value distribution of examination items (top panel) and clicker items (bottom panel). The *P* value is the fraction of students successfully solving an item, and thus the opposite of "difficulty" (thus, sometimes called "item facility"). Not surprisingly, clicker items are "easier" when considering the final answer rather than the initial one. The right panels of Fig. 2 show the point-biserial value distribution. Among the clicker items are several with negative discrimination, which fortunately do not exist among the examination items.

Cronbach's α (Cronbach 1951) is a measure of internal consistency of a test. For the examination data, the value is 0.822, which is generally considered good. For the clicker items, it is 0.774 for the initial answers and 0.79 for the final answers, which is generally considered acceptable.

Overall, based on CTT, clicker items have clearly worse psychometric properties than examination items, but the difference is not as large as one might have expected. The results are compatible with the moderate correlation that was found earlier between examination and clicker data.

Item Response Theory

CTT does not consider individual learners, e.g., whether a high-ability or a low-ability student succeeds or fails on a particular item receives equal weight; item and student properties are interwoven. Item Response Theory (IRT) on the other hand explicitly incorporates traits of the learners,

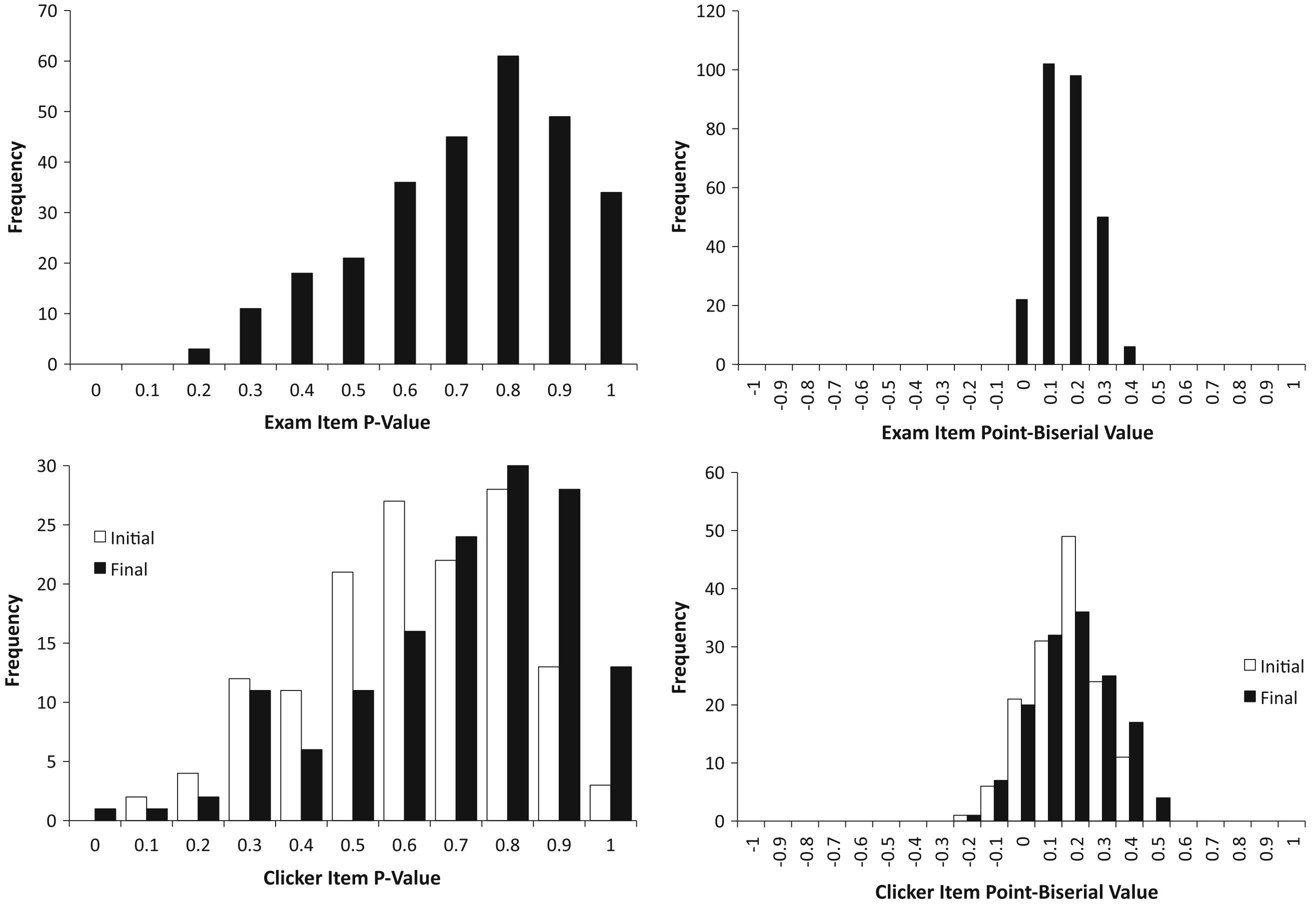


Fig. 2 Classical Test Theory item parameters for examination and clicker items in one course (*upper left panel* in Fig. 1). The *left panels* show the distribution of *P* values (item facilities), while the *right panels* show the distribution of item point-biserial values ("discriminations"). For the clicker items, estimates were based on both initial (*open*) and final (*solid*) answers

most notably "ability," and assumes that these traits influence how they interact with the problems.

IRT was originally developed in traditional examination settings (see Lord and Novick 1968 for an overview), which are highly controlled and usually high stakes. Within Physics Education Research, IRT has been used to examine the validity of concept tests (e.g., Ding and Beichner 2009; Cardamone et al. 2011), examinations (e.g., Morris et al. 2006), and online homework (e.g., Lee et al. 2008; Kortemeyer 2014).

There are a number of IRT models. The most simple model, called Rasch model, assumes that learners have one trait, their ability, and that problems have one so-called item parameter, namely their "difficulty" (Rasch 1993). The discrimination of items enters as an additional item parameter in two-parameter logistic (2PL) models (Birnbaum 1968). Beyond these commonly used models, there are also 3PL models, which incorporate guessing on a per-item base (Birnbaum 1968), as well as multidimensional IRT models, which add more learner traits in the form of additional "abilities" (Reckase 1997). However, these more complex models might overfit the data (Kortemeyer 2014).

As we are particularly interested in the discrimination, this study will use the simplest model that incorporates it, namely the 2PL model. It assumes that based on a learner j's ability θ_j, the probability for this learner j correctly answering problem i can be modeled as $p_{ij} = p_i(\theta_j)$:

$$p_{ij} = \frac{1}{1 + \exp\bigl(a_i(b_i - \theta_j)\bigr)} \ . \tag{1}$$

Here, θ_j models the ability of learner j, b_i the difficulty of item i, and a_i the discrimination of item i.

As the large number of possible models shows, this functional form is somewhat arbitrary: Essentially, Eq. 1 is one of many possible functions that have the right asymptotic properties and that have a smooth transition between "likely to not solve" and "likely to solve" that can be controlled easily by a small number of meaningful parameters. What each of the parameters does can best be illustrated using the graph of the function p_{ij}, which is known as the item characteristic curve. Figure 3 shows examples of item characteristic curves with different values of a_i and b_i. For an item with positive discrimination, a high-ability student is more likely to solve it than a low-ability student. How rapidly the probability changes with increasing ability is determined by the discrimination parameter a_i, which determines the slope at the point of inflection that is determined by the difficulty b_i. This difficulty parameter shifts the whole curve to the left or the right.

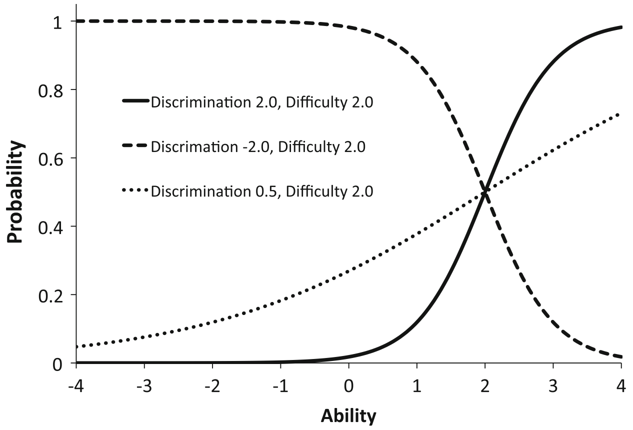


Fig. 3 Examples of item characteristic curves for different discrimination and difficulty parameters (Kortemeyer 2015). The abscissa is student ability θ_j, the ordinate the function $p_{ij} = p_i(\theta_j)$ for different a_i and b_i, which indicates the probability for a student with ability θ_j to get item i correct

Calculations are performed using the Latent Trait Model (`ltm`) package (Rizopoulos 2006) within the R statistical software system 2008. Figure 4 compares the distributions of the difficulty and discrimination parameters of examination items and clicker items in the same course. Similar to the findings using CTT, the examination item parameters have a limited range of difficulties and positive discriminations. The clicker items, on the other hand, appear to suffer from a variety of issues: Their difficulties are widely distributed, and a fraction of them have negative discrimination. The distributions are slightly better for the final than for the initial answers.

The IRT results are generally compatible with the CTT results, with a tendency to amplify the differences between the characteristics of examination and clicker data. At first glance, this result is disturbing, as it suggests that a large fraction of the clicker questions posed over the course of the semester were actually invalid assessments. To understand why the clicker items appear to be of such varying quality, it is important to investigate their properties in context. How were these items actually used within lecture sessions?

Lecture Level Measures

The lecture session under investigation is the first lecture on momentum and collisions. This lecture session is presented as an illustrative case study, and data from lecture notes, clicker software log files, and slides were used to reconstruct it. It is very typical for the lectures in the courses under investigation, as the same two instructors

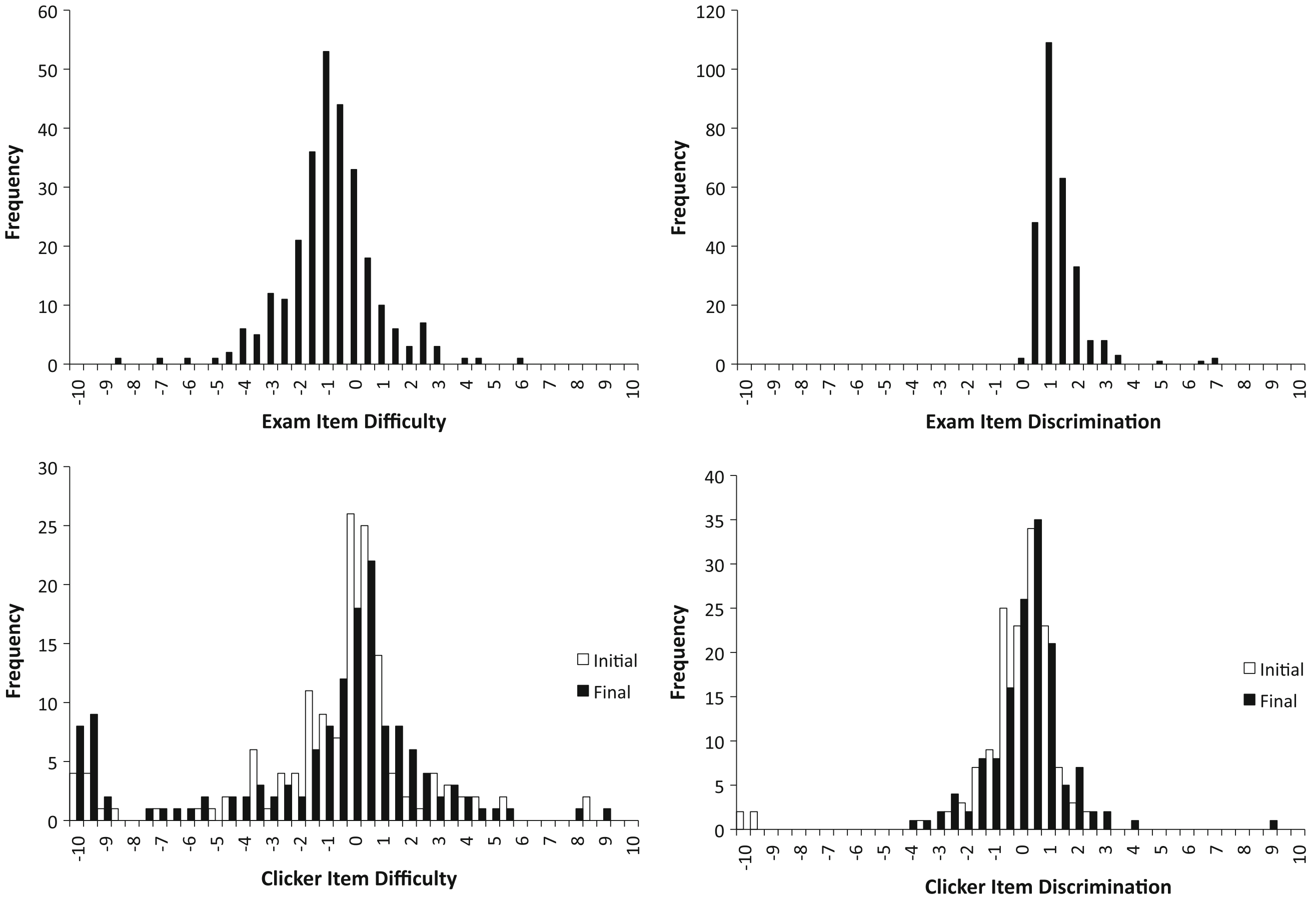


Fig. 4 Item Response Theory item parameters for examination and clicker items in one course (*upper left panel* in Fig. 1). The *left panels* show the distribution of item difficulties, while the *right panels* show the distribution of item discriminations. For the clicker items, estimates were based on both initial (*open*) and final (*solid*) answers

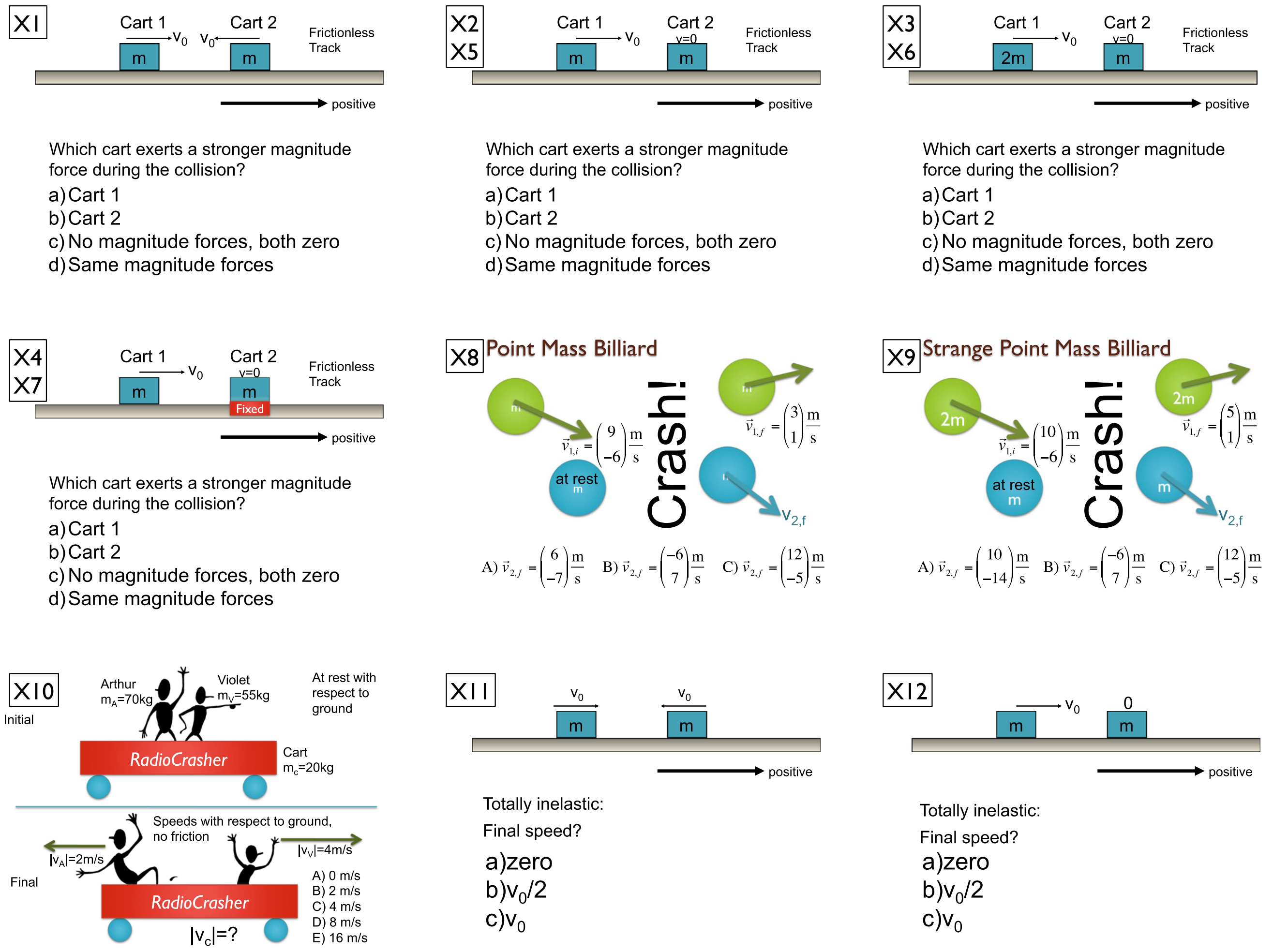
have co-taught these courses for several years in the same style.

We decided to not only consider the questions themselves, but also the context in which they were asked; consequently, the same question is treated as a separate item when it is asked again after peer discussion. We argue that this treatment is not only appropriate but necessary, since the different context in fact implicitly changes the question: The first time the question is asked, it implies, “individually consider the following question (even though we won’t stop you from talking to your neighbor),” while the second time includes the instruction “talk to and work with your neighbors, and then attempt to come to a consensus on the following question.”

The lecture was planned out by the instructor (the course has no textbook), and the questions were written by the instructor with some inspiration from Physics Education Research. However, the questions were mostly written to advance the topical coverage of the lecture session. Figure 5 shows the clicker questions asked over the course of the session, and Fig. 6 shows how these questions were embedded into the other lecture activities. Three of the questions were posed twice, before and after peer discussion, as a result of the student responses. Figure 7 shows the distribution of the student answers.

In terms of CTT, the *P* value of the items is shown in Fig. 8, and the point biserial in Fig. 9. The overall Cronbach’s α for the initial responses is 0.616 (indicating a questionable “test”), while the α for the final responses is only 0.517 (indicating a poor “test”). That Cronbach’s α decreased between initial and final responses is explained in part by the mean score increasing and the standard deviation of the score decreasing (going from 8.45 ± 2.18 for the initial response to 9.2 ± 1.83 for the final response). In either case, if this were summative assessment, the psychometrics would be alarming, and statistics like these would support the critics’ claim that clicker question results are mostly “meaningless.” However, clicker usage in this study was *not* meant to be summative, but explicitly formative.

Figure 10 shows how the overall Cronbach’s α of the “test” would change if particular items were removed;



Which cart exerts a stronger magnitude force during the collision?
a) Cart 1
b) Cart 2
c) No magnitude forces, both zero
d) Same magnitude forces

Which cart exerts a stronger magnitude force during the collision?
a) Cart 1
b) Cart 2
c) No magnitude forces, both zero
d) Same magnitude forces

Which cart exerts a stronger magnitude force during the collision?
a) Cart 1
b) Cart 2
c) No magnitude forces, both zero
d) Same magnitude forces

Which cart exerts a stronger magnitude force during the collision?
a) Cart 1
b) Cart 2
c) No magnitude forces, both zero
d) Same magnitude forces

A) $\vec{v}_{2,f} = \begin{pmatrix} 6 \\ -7 \end{pmatrix} \frac{m}{s}$ B) $\vec{v}_{2,f} = \begin{pmatrix} -6 \\ 7 \end{pmatrix} \frac{m}{s}$ C) $\vec{v}_{2,f} = \begin{pmatrix} 12 \\ -5 \end{pmatrix} \frac{m}{s}$

A) $\vec{v}_{2,f} = \begin{pmatrix} 10 \\ -14 \end{pmatrix} \frac{m}{s}$ B) $\vec{v}_{2,f} = \begin{pmatrix} -6 \\ 7 \end{pmatrix} \frac{m}{s}$ C) $\vec{v}_{2,f} = \begin{pmatrix} 12 \\ -5 \end{pmatrix} \frac{m}{s}$

A) 0 m/s
B) 2 m/s
C) 4 m/s
D) 8 m/s
E) 16 m/s

Totally inelastic:
Final speed?
a) zero
b) $v_0/2$
c) v_0

Totally inelastic:
Final speed?
a) zero
b) $v_0/2$
c) v_0

Fig. 5 Clicker items from a particular lecture. Three of the items were presented twice before and after peer discussion

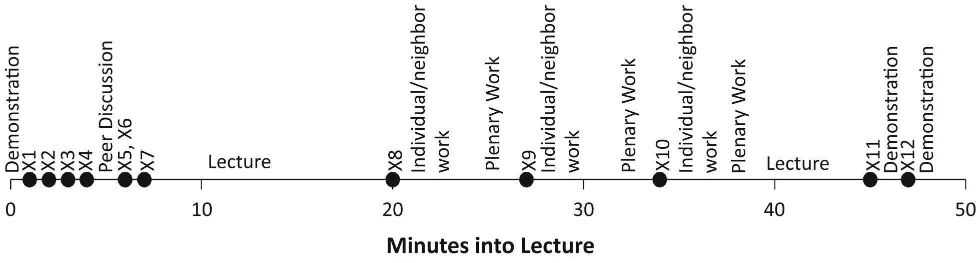


Fig. 6 Timeline of lecture activities, based on lecture notes and timestamps within the clicker log files. The *dots* indicate when the clicker items in Fig. 5 were posed

would the overall feedback from the gathered lecture data become more or less reliable if certain test items had not been posed? Items with a negative change in Cronbach's α are traditionally considered more consistent with the overall assessment than those with a positive change.

Deploying IRT, the item characteristic curves in Fig. 11) were obtained. It is obvious that the items are of varying psychometric quality. Once again, CTT and IRT results are compatible, both showing items with negative discrimination.

How could clicker assessments still be valuable? It is important to understand the interplay between the items, their function in lecture, the student responses, and the psychometric properties of the clicker data. The remainder of this section will thus walk through the lecture session and investigate each item in context.

As this was the first lecture on momentum and collisions, the session started with a short demonstration of carts colliding on an air track. The instructor commented that those will be the events that will be investigated, and

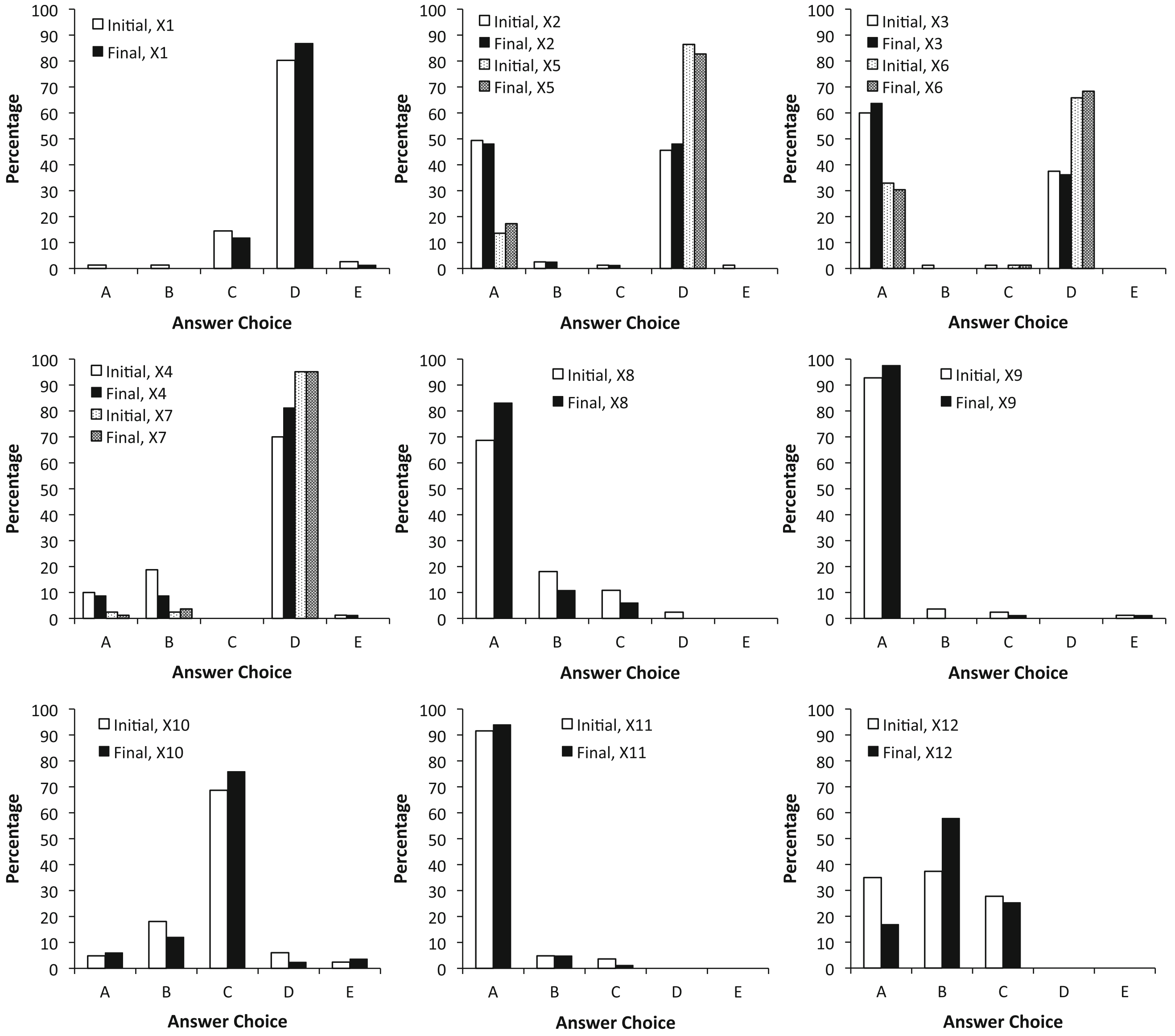


Fig. 7 Clicker item answer distributions, corresponding to the items in Fig. 5

that first one needs to understand what happens during such collisions.

Items X1 to X4

Leading up to momentum conservation, four clicker questions were posed to reiterate Newton's 3rd law. Unfortunately, some students were late arriving at lecture, and only 75, 78, 79, and 79 out of 82 students answered X1–X4, respectively.

- The first item, X1, essentially failed to elicit misconceptions, as the scenario was symmetric. The responses (Fig. 7, top left panel) indicate this: Either there are no forces, or the forces are equal. Students can get this question correct even for the wrong reasons, which can be seen from the item characteristic curve (Fig. 11, curve for item X1): The item has very low difficulty (the point of inflection corresponding to b_i is outside the plot) and very low (but still positive) discrimination (low a_i resulting in very shallow slope). Even for low-ability students, the probability of solving the item is almost 60 % for the initial response (left panel of Fig. 11). Interestingly, the discrimination increases when looking at the final answers (right panel of Fig. 11), where low-ability students only have a 40 % chance of getting it correct: It seems that low-ability students might have had second thoughts and switched

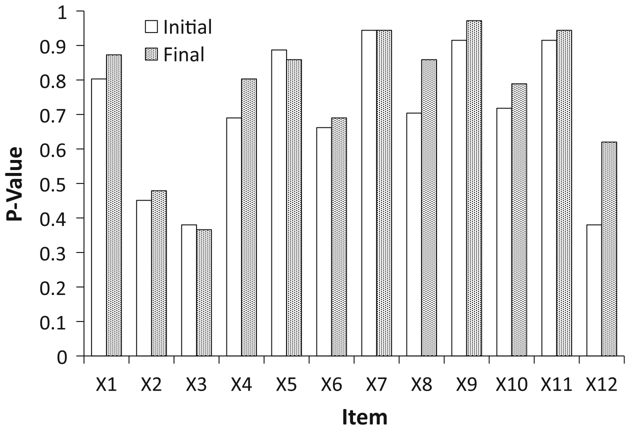


Fig. 8 *P* value (item facility) of clicker items in one particular lecture (corresponding to the items shown in Fig. 5)

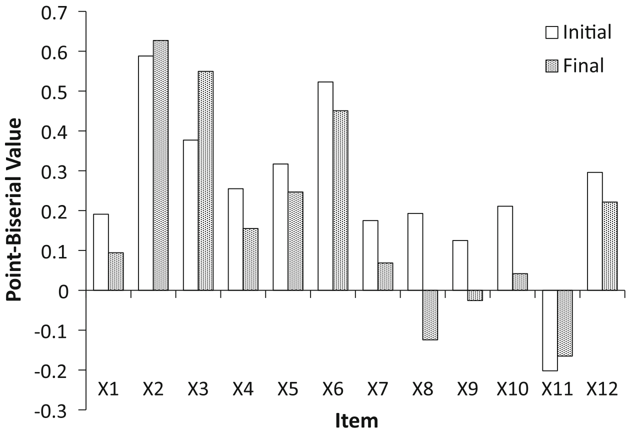


Fig. 9 Point-biserial values (item discrimination) of clicker items in one particular lecture (corresponding to the items shown in Fig. 5)

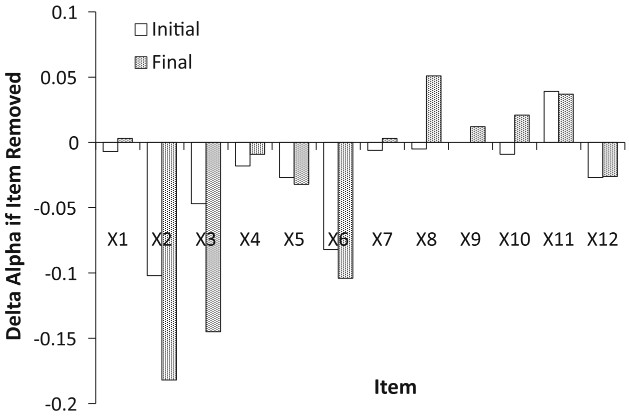


Fig. 10 Effect of removal of clicker items in one particular lecture (corresponding to the items shown in Fig. 5) on Cronbach's α. The *bars* indicate how the overall Cronbach's α would change if the indicated item was not part of the lecture

their answer from D to C, assuming that the problem just cannot be that easy, or that it is a trick question. Strangely, one student selected the non-existing option E, possibly because they came late into lecture and just clicked a random answer to get credit.

- The second item, X2, asks the same question about a non-symmetric setup. Here, misconceptions about Newton's 3rd law were clearly elicited: The student answers show half the students answering that the moving cart exerts a higher magnitude force (Fig. 7, top middle panel, first two bars).

  According to CTT, this is the best item on the "test." The point biserial is high, and if it were removed, Cronbach's α would decrease to 0.514 (from 0.616) for the initial responses and collapse to 0.335 (from 0.517) for the final responses. IRT delivers a complementary result: The item has an extremely high discrimination and cleanly distinguishes low- and high-ability students with average difficulty (Fig. 11, curve for item X2).

  The answers were shown to the students, but the instructor did not comment beyond "interesting."
- Instead of immediately discussing X2, the instructor moved on to item X3, another non-symmetric scenario, in fact doubly so, since the cart masses differ. The difference in mass drove even more students to answer that the moving and more massive cart exerts a bigger magnitude force.

  According to CTT, this is the second-best item; were it removed, α would decrease to 0.569 for the initial and 0.372 for the final responses, respectively. IRT is complementary: The item characteristic curve of X3 shows the same strong features as that for X2; high discrimination, slightly more difficult than X2.

  Once again, the instructor did not comment.
- Item X4 is the last one in the initial series of questions, and this time the non-moving cart is fixed on the track ("has the parking brake on"). Now, interestingly, answer B starts to appear: The reasoning may be that when you crash into a parked car, that car is in the way and exerts a strong force on you that makes you stop or even bounce off. However, while possibly for the wrong reasons, answer D is now in the majority: For whatever reason, both exert forces. As a result, the difficulty of the item is smaller than for X2 and X3.

At this point in the lecture, the instructor stated that there apparently has been very little agreement on the last three items, and asks the students to discuss these scenarios with each other (Peer Instruction phase). This decision was based on the clicker feedback and not planned.

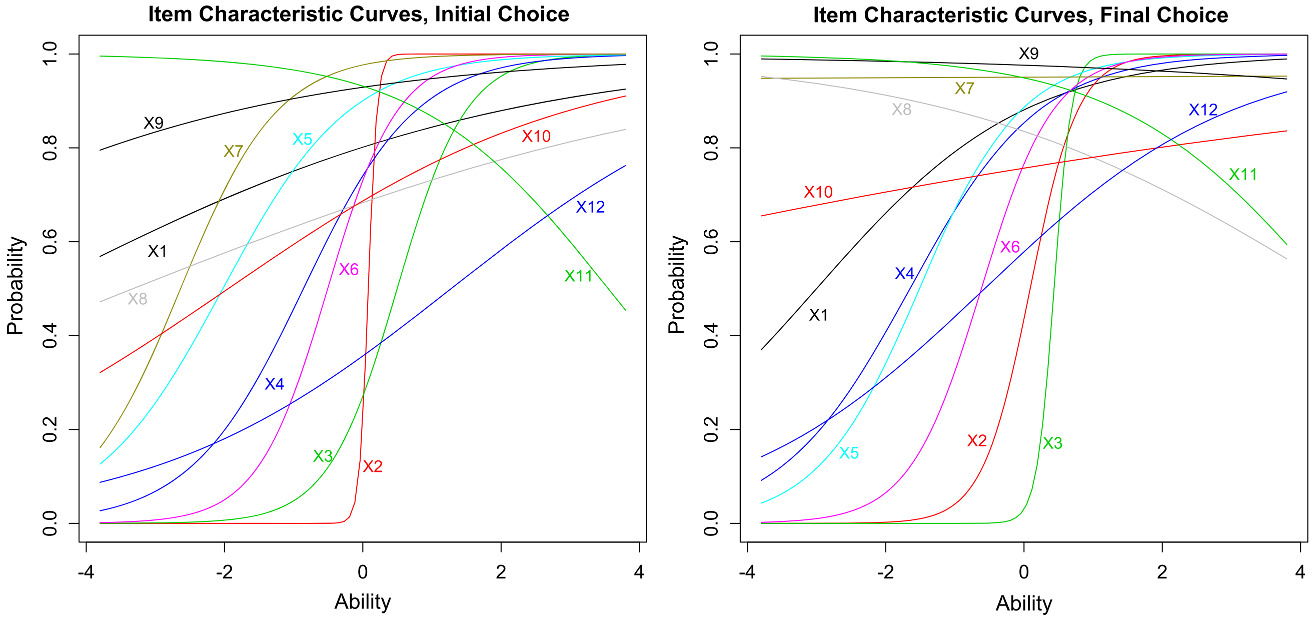


Fig. 11 Item characteristic curves for clicker items in one particular lecture, corresponding to the items shown in Fig. 5. The *left panel* is based on initial answers, while the *right panel* is based on final answers

Items X5 to X7

After the discussion calmed down, the same questions as X2 to X4 were posed again in rapid succession (after “rewinding” the slides). This time, later in the lecture, 80, 78, and 81 out of 82 responded, respectively.

- Item X5 is the same question as item X2. Clearly, after discussion, more students chose the correct answer, which was to be expected. Thus, however, the apparent difficulty of the item decreases (curve X5 in Fig. 11, compared to the curve of item X2). After peer discussion, some low-ability students might have been convinced by the arguments of high-ability students; thus, also the apparent discrimination of the item decreases.
- Item X6 is the same question as item X3. Also here, the correct answer was chosen more frequently; however, answer A is still a strong contender, reflecting the residual preference for this choice during the first round. While also here, difficulty and discrimination decreased compared to X3, the item is still a strong question.
- Item X7 is the same question as item X4. Something very interesting happened here, as peer discussion mostly eliminated choices A and B. Apparently, the arguments why one or the other force should be stronger became untenable compared to simply saying that the forces are equal. Thus, the apparent difficulty and apparent discrimination dramatically decreased (almost flat curve X7 in Fig. 11).

Thus, the apparent quality of items X5, X6, and X7 is lower than when the questions were asked the first time around, while in fact, the students learned during the intervening period of Peer Instruction; context is important. Based on Fig. 10, however, with the exception of X7, omitting these items from the lecture would have led to less consistent feedback. In other words, even though Peer Instruction moved the assessment from individual to collective performance, the gathered feedback is still meaningful. This same desirable consistency is not achieved by the following group of numerical items, where students were encouraged to work in groups from the get-go.

Items X8 to X10

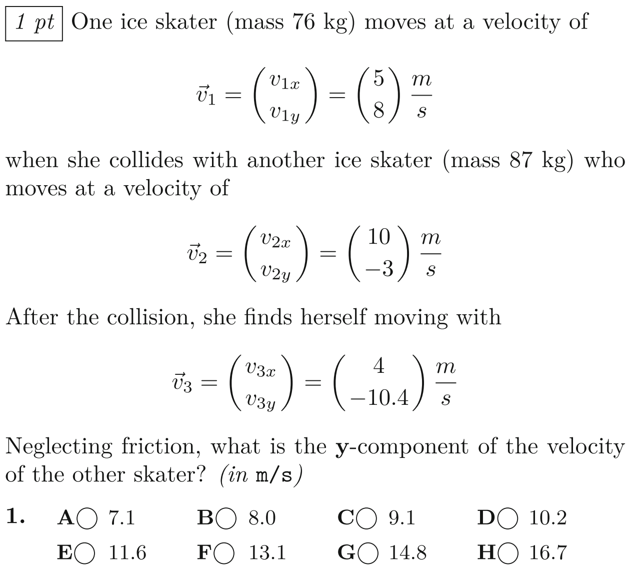
The session continued with a 15-min lecture segment discussing the implications of Newton’s 3rd law and deriving momentum conservation. The next set of three questions was designed to practice calculations involving momentum conservation. Students were asked to calculate the results and encouraged to talk and check in with their neighbors. All 82 students were responding. Following the clicker votes, the calculations were reiterated as a plenary presentation and discussion.

- Item X8 is a simple calculation problem. Choices A and B are in opposite directions (indicating a sign error), while choice C would result if a student added the initial and final momentum.
  Interestingly, removal of this item would slightly decrease Cronbach’s α of the initial responses, but

increase α of the final responses. IRT shows a complementary result: The discrimination changes from slightly positive to slightly negative between the initial and final responses (curve X8 in Fig. 11). It is not clear why the discrimination is negative, but it may be possible that high-ability students did not bother to calculate this simple but work-intensive problem, and instead just submitted a random answer. The problem also has very low difficulty.

The final examination for this course included a similar problem, see Fig. 12. As it turns out, the Phi coefficient of association between clicker and examination correctness is slightly positive for the initial clicker response ($r_\phi = 0.14$), but negative for the final clicker response ($r_\phi = -0.21$), compatible with the above results that indicate that for this question, the initial answer is more meaningful than the final answer. In either case, the correlation is very weak, and the clicker question would be an unsuitable predictor for examination performance.

- Item X9 was more complicated than item X8, as the masses were different. However, more students got this item correct, presumably because the calculations for item X8 had been demonstrated. Thus, the apparent difficulty is even lower than that of X8, and the discrimination is almost zero.
- Item X10 is a more complicated problem, in that it involved three bodies and a possible confusion about the frame of reference. While the majority of students got this problem correct, random other choices are also made. The difficulty is higher than for X8 and X9, but the discrimination is low.



$\boxed{1\ pt}$ One ice skater (mass 76 kg) moves at a velocity of

$$\vec{v}_1 = \begin{pmatrix} v_{1x} \\ v_{1y} \end{pmatrix} = \begin{pmatrix} 5 \\ 8 \end{pmatrix} \frac{m}{s}$$

when she collides with another ice skater (mass 87 kg) who moves at a velocity of

$$\vec{v}_2 = \begin{pmatrix} v_{2x} \\ v_{2y} \end{pmatrix} = \begin{pmatrix} 10 \\ -3 \end{pmatrix} \frac{m}{s}$$

After the collision, she finds herself moving with

$$\vec{v}_3 = \begin{pmatrix} v_{3x} \\ v_{3y} \end{pmatrix} = \begin{pmatrix} 4 \\ -10.4 \end{pmatrix} \frac{m}{s}$$

Neglecting friction, what is the **y**-component of the velocity of the other skater? *(in* `m/s`*)*

1. **A**○ 7.1 **B**○ 8.0 **C**○ 9.1 **D**○ 10.2
E○ 11.6 **F**○ 13.1 **G**○ 14.8 **H**○ 16.7

Fig. 12 An examination problem similar to clicker item X8

Both the responses and the item parameters indicate that these calculation problems are not meaningful. Here, clickers merely provided an incentive to actually do the calculations, but the feedback gathered is essentially useless in terms of providing formative assessment. It is however revealing that these calculations were so much easier than the “simple” questions that could be answered based solely on the understanding of Newton’s 3rd law.

Items X11 and X12

Totally inelastic collisions were introduced during a short lecture segment, followed by two more clicker questions. All 82 students were responding. In these conceptual questions, the students were asked to predict the outcome of two experiments on the air track. After the voting was closed, the instructor “let nature decide” which answer was correct.

- Item X11 shows surprising properties. According to CTT, Cronbach’s α increases if this item is removed, namely to 0.655 (from 0.616) for the initial and 0.554 (from 0.517) for the final answers. The point biserial is negative. According to IRT, while most students arrived at the correct solution (which is also reflected by $p_{ij} \approx 1$ for low-ability students), the discrimination is negative. It is possible that high-ability students were overthinking the problem, assuming that it just cannot be that simple.
- Item X12 again shows positive discrimination in spite of the answer choice being more randomly distributed. High-ability students may have realized that totally inelastic collisions are indeed very simple, or the previous demonstration of seeing what an inelastic collision looks like may have helped.

The instructor had more questions of this type prepared, but ran out of time due to the earlier second round on the Newton’s 3rd law questions.

Limitations and Discussion

This case study was carried out in a course setting with a particular population, grading method, and educational philosophy (see Sect. 2), and thus results are not necessarily universally applicable. Having a majority of premedical students and giving credit for answering questions resulted in an extremely high level of student participation, and having a slight grading advantage for answering questions correctly cut down on random answers. The exact effect of these factors would need to be investigated using data from other institutions, but it can be surmised that the results in this study represent the upper limit of "meaningfulness."

Last but not least, the study is limited to the subject matter of physics, which arguably was the first field to widely apply clickers in higher education settings and thus has the longest tradition of best practices. Usage of clickers in other fields, even within other natural sciences, may result in different outcomes.

Conclusions

Classroom response systems should generally not be used for testing purposes; instead, they are a tool to foster learning. Nevertheless, it is worthwhile to investigate the psychometric properties of these questions. In this case study, using global correlations with examination data, as well as methods of both Classical Test Theory and Item Response Theory, it turns out that the psychometric properties of clicker data, while worse than those of examination data, still provide meaningful feedback:

- Clicker data are a moderate predictor of examination performance.
- The internal consistency of clicker data is acceptable on the scope of a complete course.
- The difficulty of clicker items is more widely distributed than the difficulty of examination items, but generally comparable.
- The discrimination of clicker items is also more widely distributed than the discrimination of examination data, and while there are a number of items with negative discrimination, there are also items that have larger discrimination than any examination items.
- In a case study of a particular lecture session, the psychometric properties of clicker items could be explained in terms of the involved physics and the function of the item within the lecture.

With regard to the latter finding, it is thus important to remember that the assessment occurs in context of lecture sessions, and thus in addition to the pure psychometric properties of the items, it is relevant when, where, and how these items are administered. It was shown that the same item can have very different apparent properties depending on which function it served and when it was posed; most notably, items change properties before and after peer discussions. In the session under investigation, the responses and item properties provided highly meaningful situational feedback, and modifying lecture pace and topical coverage based on this feedback was appropriate. A possible exception was questions involving calculations, but here the questions also served a different purpose, mainly just to hold the students accountable to actually do the exercises.

Overall, even the non-scientifically constructed clicker questions used in this case study provided meaningful feedback: moderately so in terms of psychometrics, but definitely so in terms of useful feedback during lectures. While planning the lecture, simply finding places where questions may be appropriate or useful, and then creating questions that fit in that particular context appears to be a justifiable approach; within reasonable limits, the context and educational function may be even more important than having the "perfect" question. There will even be blatantly imperfect questions, such as those which turned out to have negative discrimination, but even those might be "teachable moments"—as opposed to examination items, clicker questions are posed in an interactive and dynamic context, and instructors and students will notice that something is "wrong" and work out what happened. This approach certainly lowers the barrier to implementing this activating teaching strategy, while at the same time, not losing validity.

Acknowledgments The author would like to thank Emre Gönülateş for his helpful feedback, discussions, and suggestions.

References

Barth-Cohen LA, Smith MK, Capps DK, Lewin JD, Shemwell JT, Stetzer MR (2016) What are middle school students talking about during clicker questions? Characterizing small-group conversations mediated by classroom response systems. J Sci Educ Technol 25(1):50–61

Beatty I, Gerace W (2009) Technology-enhanced formative assessment: a research-based pedagogy for teaching science with classroom response technology. J Sci Educ Technol 18(2):146–162. doi:10.1007/s10956-008-9140-4

Birnbaum A (1968) in Lord and Novick (1968), chap. Some latent trait models and their use in inferring an examinees ability, pp 374–472

Caldwell JE (2007) Clickers in the large classroom: current research and best-practice tips. CBE Life Sci Educ 6(1):9–20

Cardamone CN, Abbott JE, Rayyan S, Seaton DT, Pawl A, Pritchard DE (2011) Item response theory analysis of the mechanics baseline test. In: Physics education research conference 2011 AIP conference proceedings, pp 135–138

Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16(3):297–334

Crouch CH, Mazur E (2001) Peer instruction: ten years of experience and results. Am J Phys 69(9):970–977

Dancy M, Henderson C (2010) Pedagogical practices and instructional change of physics faculty. Am J Phys 78(10):1056–1063

Ding L, Reay NW, Lee A, Bao L (2009) Are we asking the right questions? Validating clicker question sequences by student interviews. Am J Phys 77(7):643–650

Ding L, Beichner R (2009) Approaches to data analysis of multiple-choice questions. Phys Rev ST Phys Educ Res 5:020103. doi:10.1103/PhysRevSTPER.5.020103

Fagen AP, Crouch CH, Mazur E (2002) Peer instruction: results from a range of classrooms. Phys Teach 40(4):206–209

Henderson C, Dancy M, Niewiadomska-Bugaj M (2012) Use of research-based instructional strategies in introductory physics: where do faculty leave the innovation-decision process? Phys Rev Spec Top Phys Educ Res 8(2):020104

iClicker Classroom Response System (2003) https://www.iclicker.com. Accessed Mar 2016

James MC (2006) The effect of grading incentive on student discourse in Peer instruction. Am J Phys 74(8):689–691

James MC, Willoughby S (2011) Listening to student conversations during clicker questions: what you have not heard might surprise you!. Am J Phys 79(1):123–132

Kay RH, LeSage A (2009) Examining the benefits and challenges of using audience response systems: a review of the literature. Comput Educ 53(3):819–827

Keller C, Finkelstein N, Perkins K, Pollock S, Turpen C, Dubson M (2007) Research-based practices for effective clicker use. In: 2007 physics education research conference, vol 951, pp 128–131

King DB, Joshi S (2008) Gender differences in the use and effectiveness of personal response devices. J Sci Educ Technol 17(6):544–552

Kortemeyer G (2015) Scalable continual quality control of formative assessment items in an educational digital library: an empirical study. Int J Digit Libr 1–13. doi:10.1007/s00799-015-0145-3

Kortemeyer G, Kashy E, Benenson W, Bauer W (2008) Experiences using the open-source learning content management and assessment system LON-CAPA in introductory physics courses. Am J Phys 76:438–444

Kortemeyer G (2014) Extending item response theory to online homework. Phys Rev ST Phys Educ Res 10:010118. doi:10.1103/PhysRevSTPER.10.010118

Lantz M (2010) The use of 'clickers' in the classroom: teaching innovation or merely an amusing novelty? Comput Hum Behav 26(4):556

Lasry N (2008) Clickers or flashcards: is there really a difference? Phys Teach 46(4):242–244

Lasry N, Mazur E, Watkins J (2008) Peer instruction: from Harvard to the two-year college. Am J Phys 76(11):1066–1069

Lee YJ, Palazzo DJ, Warnakulasooriya R, Pritchard DE (2008) Patterns, correlates, and reduction of homework copying. Phys Rev ST Phys Educ Res 4:010102. doi:10.1103/PhysRevSTPER.4.010102

Lord FM, Novick MR (eds) (1968) Statistical theories of mental test scores. Addison-Wesley, Reading

Mazur E (1997) Peer instruction. Prentice Hall, Upper Saddle River

Morris GA, Branum-Martin L, Harshman N, Baker SD, Mazur E, Dutta S, Mzoughi T, McCauley V (2006) Testing the test: item response curves and test quality. Am J Phys 74(5):449–453

Nunnally J, Bernstein I (1994) Psychometric theory. McGraw-Hill, New York

Poole D (2012) The impact of anonymous and assigned use of student response systems on student achievement. J Interact Learn Res 23(2):101–112

R Development Core Team (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org. ISBN 3-900051-07-0

Rasch G (1993) Probabilistic models for some intelligence and attainment tests. MESA Press, Chicago

Reckase MD (1997) The past and future of multidimensional item response theory. Appl Psychol Meas 21(1):25–36

Richardson AM, Dunn PK, McDonald C, Oprescu F (2014) CRiSP: an instrument for assessing student perceptions of classroom response systems. J Sci Educ Technol 24(4):432–447

Richardson CT, O'Shea BW (2013) Assessing gender differences in response system questions for an introductory physics course. Am J Phys 81(3):231–236

Rizopoulos D (2006) Ltm: an R package for Latent Variable Modelling and Item Response Theory Analyses. J Stat Softw 17(5):1–25. http://www.jstatsoft.org/v17/i05/

Setzer JC, Wise SL, van den Heuvel JR, Ling G (2013) An investigation of examinee test-taking effort on a large-scale assessment. Appl Meas Educ 26(1):34–49

Turpen C, Finkelstein ND (2009) Not all interactive engagement is the same: variations in physics professors implementation of Peer Instruction. Phys Rev Spec Top Phys Educ Res 5(2):020101

White P, Syncox D, Alters B (2011) Clicking for grades? Really? Investigating the use of clickers for awarding grade-points in post-secondary education. J Interact Learn Environ 19(5):551–561

Willse JT (2014) CTT: Classical Test Theory Functions. https://cran.r-project.org/web/packages/CTT/. Accessed Mar 2016