frontiers in Education

ORIGINAL RESEARCH
published: 31 August 2021
doi: 10.3389/feduc.2021.735088

Setting Standards With Multiple-Choice Tests: A Preliminary Intended-User Evaluation of SmartStandardSet

Gavin T. L. Brown[1]*, *Paul Denny*[2], *David L. San Jose*[3] and *Ellen Li*[4]

[1]Faculty of Education and Social Work, The University of Auckland, Auckland, New Zealand, [2]Faculty of Science, School of Computer Science, The University of Auckland, Auckland, New Zealand, [3]Faculty of Education and Social Work, The University of Auckland, Auckland, New Zealand, [4]Department of Computer Science, Faculty of Science, The University of Auckland, Auckland, New Zealand

Software that easily helps higher education instructors to remove poor quality items and set appropriate grade boundaries is generally lacking. To address these challenges, the SmartStandardSet system provides a graphical-user interface for removing defective items, weighting student scores using a two-parameter model IRT score analysis, and a mechanism for standard-setting. We evaluated the system through a series of six interviews with teachers and six focus groups involving 19 students to understand how key stakeholders would view the use of the tool in practice. Generally, both groups of participants reported high levels of feasibility, accuracy, and utility in SmartStandardSet's statistical scoring of items and score calculation for test-takers. Teachers indicated the data displays would help them improve future test items; students indicated the system would be fairer and would motivate greater effort on more difficult test items. However, both groups had concerns about implementing the system without institutional policy endorsement. Students specifically were concerned that academics may set grade boundaries on arbitrary and invalid grounds. Our results provide useful insights into the perceived benefits of using the tool for standard setting, and suggest concrete next steps for gaining wider acceptance that will be the focus of future work.

Keywords: data science applications in education, human-computer interface, pedagogical issues, post-secondary education, teaching/learning strategies

OPEN ACCESS

Edited by:
Christopher Charles Deneen, The University of Melbourne, Australia

Reviewed by:
Katrien Struyven, University of Hasselt, Belgium
Jeremy R. Sullivan, University of Texas at San Antonio, United States

***Correspondence:**
Gavin T. L. Brown
gt.brown@auckland.ac.nz

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *02 July 2021*
Accepted: *17 August 2021*
Published: *31 August 2021*

Citation:
Brown GTL, Denny P, San Jose DL and Li E (2021) Setting Standards With Multiple-Choice Tests: A Preliminary Intended-User Evaluation of SmartStandardSet. Front. Educ. 6:735088. doi: 10.3389/feduc.2021.735088

HIGHLIGHTS

- SmartStandardSet provides industry standard mechanisms to evaluate MCQ item quality.
- Instructors can use the system to evaluate the quality and difficulty of their own tests, leading to improved future test writing.
- SmartStandardSet allows instructors to set appropriate grade-standards when tests are deemed too easy or too hard.
- SmartStandardSet has the potential to improve item writing and student learning behaviours.
- For institutions of higher education, deployment of SmartStandardSet can ensure quality in standardised testing practices.

INTRODUCTION

Grade boundaries for tests are usually related to the proportion of items answered correctly. This is potentially misleading because test difficulty or easiness is not considered (e.g., easy tests create high scores). This is partially resolved if test scores are created with Item Response Theory (IRT) methods that create scores adjusted by question difficulty, rather than counting answers correct. Standard setting protocols have been developed to convert test scores into appropriate letter grades. However, both IRT and standard setting protocols are time-consuming and labour-intensive processes and thus infrequently implemented in higher education. Moreover, there may be resistance from both students and lecturers in accepting IRT-based scoring in environments where it is not approved by policy. This paper addresses both gaps by: 1) describing a newly developed prototype tool, SmartStandardSet, for performing test quality evaluation and standard setting, and 2) conducting an exploratory pilot evaluation from the perspective of intended stakeholders concerning the utility, feasibility, accuracy, and propriety of the system. This preliminary evaluation gauges the acceptance of a potentially major change in how multiple-choice tests are evaluated and prepared for grading; hence, it is warranted and provides useful insights despite its small-scale.

Test Scoring and Standard Setting

The use of classical test theory (CTT) in scoring multiple-choice question (MCQs) tests is widely adopted in higher education. In many influential disciplines (e.g., medicine, sciences, engineering) they are used as a reliable and efficient means of operationalising large-scale, summative evaluation of student learning and content knowledge (McCoubrie, 2004; Butler, 2018; Joshi et al., 2019). We recognize that there is a general impression that MCQs only require recall and recognition of correct answers and thus have limited validity. There are grounds for considering that MCQs can test analysis, relational thinking, and even abstract reasoning if constructed appropriately (Hattie and Purdie, 1998; Melser et al., 2020). Indeed direct comparison of the use of MCQs and constructed response questions on examinations has shown that, in some contexts, statistically similar results can be generated from each question type (Ventouras et al., 2010; Herzog et al., 2019). Regardless of these comparisons, MCQs remain a popular format because they allow breadth of coverage within a content domain, efficiency in administration and scoring, and contribute to learning.

MCQ tests are generally scored, in accordance with CTT, as the sum of the number of items answered correctly, without any weighting for more difficult or easier items. Further, few institutions provide or require test item analysis to identify and remove psychometrically poor items (e.g., negative discrimination or extreme difficulty) before test scores are finalised. The percentage correct scores are conventionally mapped to letter grade ranges (e.g., $A \geq 90\%$ or $50\% \leq C \leq 65\%$, etc.) as if there is no uncertainty as to the validity of such mapping given the varying difficulty of tests. Essentially, computing grades this way is cheap, simple to do, easy to understand, and efficient. However, this reduces letter grades to indicators of quantity rather than signals of quality.

This situation creates two major problems for quality assurance of grades in higher education. The first, rather more technical, lies in item analysis methods and procedures to eliminate defective items and weight scores according to item difficulty. IRT item analysis permits identification of poor-quality items and the adjustment of scores weighted by the difficulty of items answered correctly (Hambleton et al., 1991; Embretson and Reise, 2000; Schauber and Hecht, 2020). This is necessary because faculty-written mid-term and final MCQ examinations can be of very poor quality (Brown and Abdulnabi, 2017). However, the software used to estimate these parameters (e.g., R packages mirt or ltm; WinSteps; bilog, etc.) are not easy to use for the non-psychometrician end-user. The irony of this situation is that standardized university admissions tests tend to use IRT methods to analyse items and generate scores, while university testing itself is largely CTT or judgement-based (Halpern and Butler, 2013). The situation in operational testing in higher education is that statistical item analysis has to be conducted retrospectively (i.e., after the test is administered) rather than prospectively as would be the case in item banking (Wright and Bell, 1984). Administrative requirements, which result in exposure of previous examinations to students, mean that every examination has to be new to reduce the possibility of cheating. This means that the use of pre-calibrated item banks is problematic. Hence, mechanisms for analysing rapidly the quality of instructor-created tests are necessary.

The second challenge lies in the need for content-expert professional judgment to map test difficulty to qualities of performance (e.g., excellent, good, satisfactory, or unsatisfactory). A wide variety of standard setting procedures exist (Cizek, 2001; Blömeke and Gustafsson, 2017; Afrashteh, 2021) by which experts can set defensible cut-scores to reflect increments in quality once item difficulties are established empirically with IRT analysis. One such procedure is the bookmark method (Mitzel et al., 2001; Baldwin et al., 2020) in which experts insert "bookmarks" on items that represent increments in quality categories. While conceptually simple, the protocols require some administrative infrastructure which could be readily digitized. Nonetheless, while standard setting decisions should be made independent of the person-score distribution, Angoff (1974) made it clear that "if you scratch a criterion-referenced interpretation, you will very likely find a norm-referenced set of assumptions underneath" (p. 4). This means that instructors are highly likely to consider both standards relative to item content/difficulty and the distribution of test-takers when setting cut-scores for grade standards.

Technology Acceptance

Technology acceptance situates technology use within a theory of planned or reasoned behaviour (Ajzen, 1991). The assumption is that user behaviour depends on positive attitudes and beliefs and a sense of control around technology. Technology acceptance research in education has suggested that provision of technology alone cannot guarantee the effective and expected use of

technology (Teo, 2010; Scherer et al., 2019). Hence, it is important to understand the perspective of intended end users, who, in this context, are university instructors and students. Their perceptions relate to the end-goal of summative MCQ testing, which is to generate sufficiently reliable performance information to allow valid grades to be awarded to test-takers. Because assessment technologies might not impact all stakeholders (e.g., instructors and students) equally, it is important to move the focus of research from technology functionality to user perceptions and experiences (Katz and Gorin, 2016).

In the context of introducing computer-assisted testing into the compulsory school sector (Hattie, et al., 2006), teacher beliefs about the purpose and nature of assessment mattered to successful interpretation of the software. While students and teachers can be enthusiastic about novel technologies in education, this tends to be the case when consequences attached to performance are low to zero. Students tend to be resistant toward assessment innovations that count toward grades, perhaps because their previous success was based on traditional or conventional assessment practices (Struyven and Devesa, 2016). Thus, the validity of technology use for testing purposes depends on positive beliefs by potential end users as to the technology's capacity to 1) support accurate performance scores, 2) provide utility for instructors and student test-takers, and 3) meet ethical and regulatory expectations and constraints. Explicit attention to maximizing end-user beneficial consequences (Sen, 2000), rather than on statistical or visual elegance needs to be the objective in evaluating new technologies.

Research Questions

In this paper, we describe a software system, SmartStandardSet, that automates IRT analysis of MCQ test items, calculates weighted scores for students, and allows for grade boundaries to be easily set according to standards-based judgements by higher education instructors. Second, and more importantly, we report a contextual evaluation to identify whether implementation of the software would achieve the intended quality of experience and benefits so as to justify its wider deployment and usage in practice. Through this evaluation, we address the following two research questions:

RQ1) What concerns do university lecturers have, and what potential benefits do they see, in adopting IRT-based scoring and standard setting protocols in educational contexts?

RQ2) To what extent do university students accept grading standards that have been set through a combination of IRT test-scoring and instructor judgment in high-stakes examinations?

SMARTSTANDARDSET: AN EDUCATIONAL TECHNOLOGY TESTING INNOVATION

SmartStandardSet is a new, user-friendly, web-based tool that implements the two-parameter model (2 PL) IRT method in the R package "mirt" (Chalmers, 2012). SmartStandardSet automatically identifies psychometrically invalid items (i.e., negative discrimination indices, all correct, or all wrong). The status of each item is signaled by a red cross or green tick to indicate invalid or valid, respectively (see **Figure 1**). Optionally, users may include invalid items or even remove valid items with relatively flat but positive discrimination slopes. In the example in **Figure 1**, item 4 has been plausibly selected by the user for exclusion perhaps because of its extreme ease (b = −5.04, percent correct = 95.7%) and weak positive discrimination (a = 0.66). Although SmartStandardSet automatically flags invalid items for removal, the instructor remains in full control and can choose to override such recommendations if they wish. Such decisions may especially arise in the case of items that 100% of test-takers got correct. While such easy items do not contribute to discrimination, removal may contribute to loss of morale and motivation among test takers; from their perspective answering easy items correctly should contribute to their test performance because the success of others ought not to impact their own score.

The time required for SmartStandardSet to compute discrimination and item weight values depends on the file size (i.e., the number of items and number of test-takers), but is generally a matter of seconds, a consequence of how the package "mirt" has been optimised. For example, input files containing responses to an 80-question test from 1,200 students, and an answer key for those 80 questions, are processed in approximately 10–15 s through the web-based interface.

Once the user has decided on the items to be kept, SmartStandardSet creates a 2 PL IRT score for each test-taker, on a common-item design basis (i.e., all test-takers take all items; Kolen, 2006), and displays the items in a custom-created Wright Map (Wright and Stone, 1979), (see **Figure 2**). The Wright Map orders items and people based on performance and difficulty. Scores for both people and items are indexed with IRT in a range, normally from -3 to +3, and have $M = 0$, and $SD = 1$. Normally, the items are ordered from easiest at the bottom of the chart to hardest at the top. Likewise, students are ordered from least proficient at the bottom of the chart to most proficient at the top. In SmartStandardSet, items are displayed on the left in red, while people are displayed on the right in blue. Ideally, all person ability scores will be evaluated with sufficient test items with matching difficulties to produce small error ranges. **Figure 2** shows a large number of people are in the score range ≥+2.00 but that there are no test items in the same range. Conversely, six test items are easier than −2.00, without any people in the same difficulty range. This could suggest that in future tests, harder items are needed, for which space could be created in the test by removing the very easy items. At any time, decision makers can see how the distribution of items and their cut-scores relate to the distribution of test-takers which is displayed or hidden on the right side of the Wright Map.

When SmartStandardSet first opens this page, the distribution of students is obscured so that only the item distribution is seen. Ideally, instructors set grade boundaries using item content and item difficulty location; they indicate this by moving a slider to reflect their judgment of where four major grades (i.e., D/C = fail, unsatisfactory vs. pass, satisfactory; C/B = pass, satisfactory vs. pass, good; and B/A = pass, good vs. pass, excellent) begin and end. **Figure 3** shows that the boundaries for the beginning of grades C, B, and A are set at −3.00, −1.50, and −0.80, respectively.

Number	Question data	% Correct	Discrimination	Item weight	Keep / Discard
1		84.2	1.5	-1.5	✓
2		86.9	1.23	-1.91	✓
3		83.9	0.65	-2.76	✓
4		95.7	0.66	-5.04	✗
5		97.9	1.89	-2.87	✓
6		83.7	1.13	-1.75	✓

FIGURE 1 | SmartStandardSet 2 PL IRT Item Analysis Display (the user has manually selected to discard item 4).

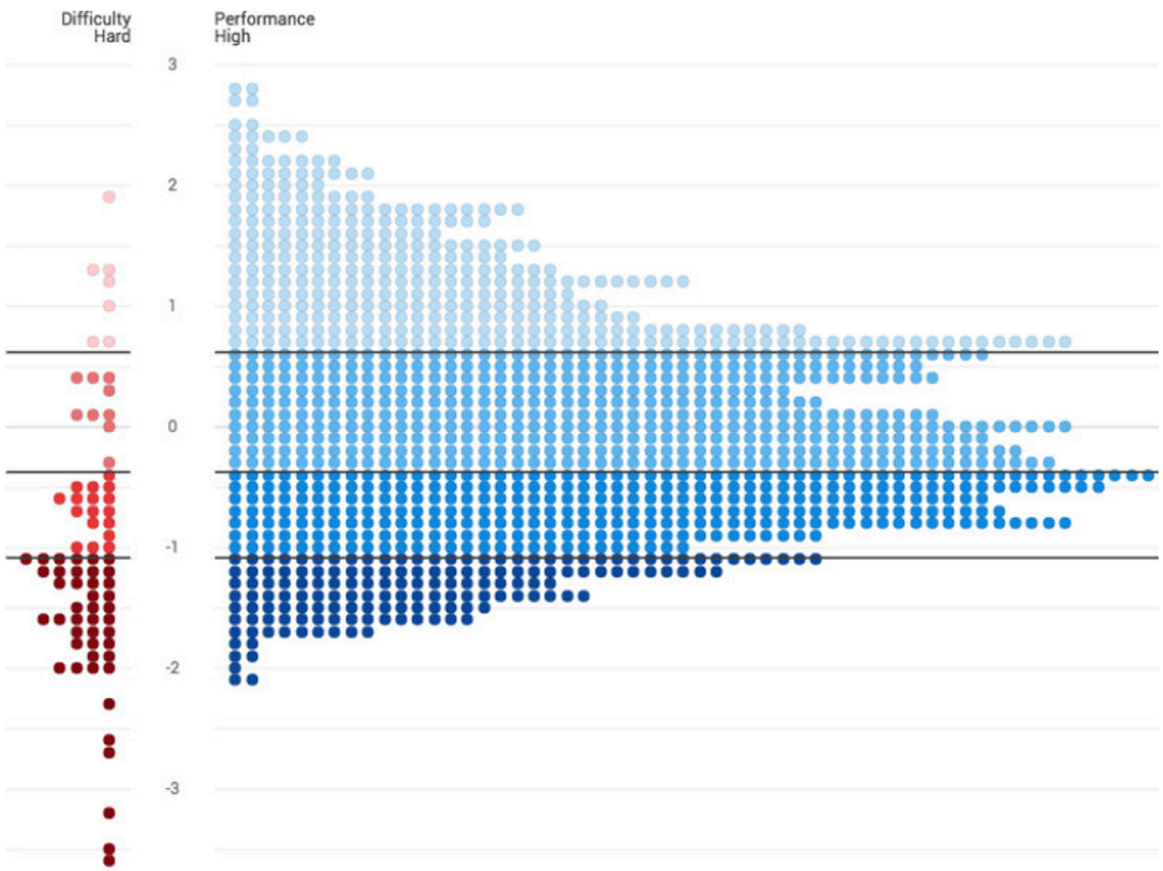


FIGURE 2 | SmartStandardSet IRT and person score location display.

As each boundary is set, items and people indicators change to a different shade of red and blue, respectively.

Although SmartStandardSet makes an initial estimate of where these cut scores could lie, by averaging the positions of test-takers who achieve raw scores equal to commonly used grade boundaries (i.e., 50% raw score in CTT equates to the start of Grade C), the instructor is entirely in control of where to place each boundary. This standard-setting process continues with sub-grade cut scores (i.e., + and–for each letter grade) set for the plus and minus units of each grade. **Figure 4** shows that boundaries for D+ and D-are spaced such that one student falls in D-, one in D, and one in D+, and three on the cusp of D+ and C-. The cut

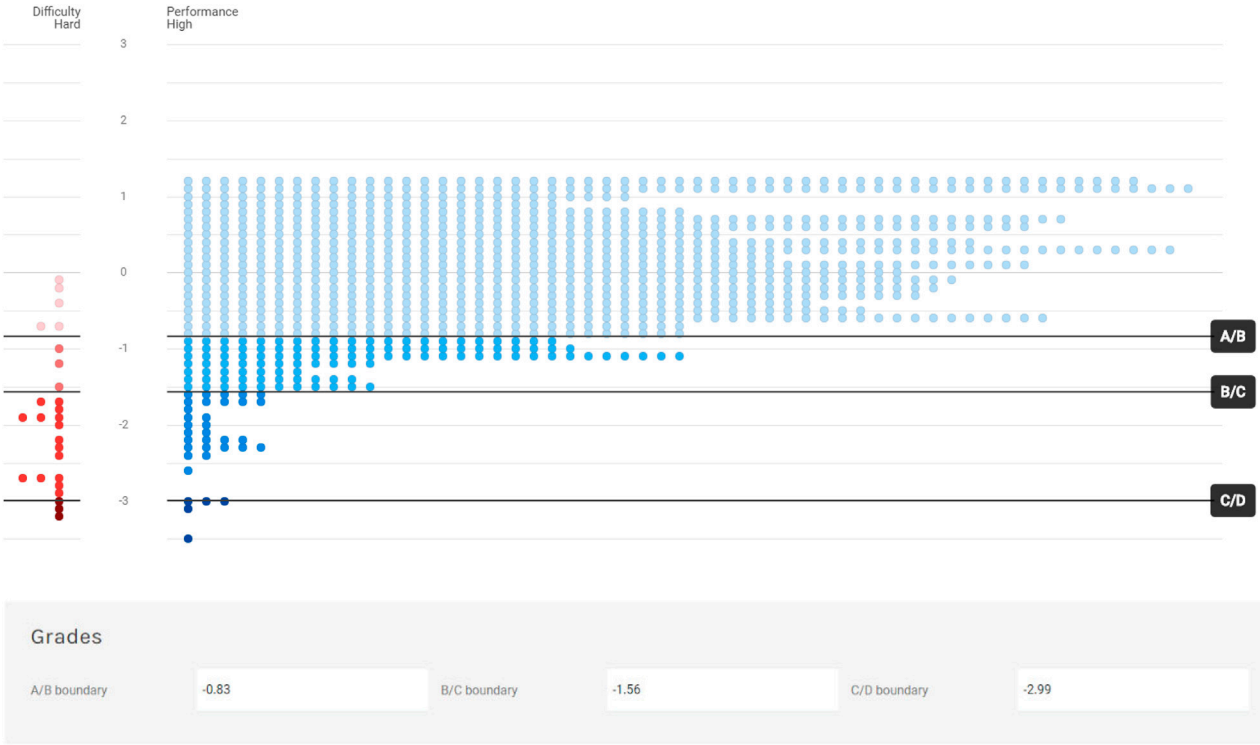


FIGURE 3 | SmartStandardSet standard setting display: Major grade boundaries.

scores for C- and C+ are spaced equally between the C/D and B/C cut scores. The A range cut scores fall such that one-third of all A students get A-, one-third get A, and a third get A+. Remember these automated settings can be over-ridden by the instructor by dragging the boundary handles.

Once the grade boundaries are set, SmartStandardSet transforms the student IRT scores from the logit scale to the university required percentage scores and grades (see **Figure 5**). This successfully transforms the IRT score centered on zero, to a meaningful scale within the institution. Additionally, a summary of changes in student scores in both grades and examination scores as a consequence of item removal or boundary setting is provided. **Figure 5** shows that in the A range, 55 students moved upward by two or more sub-grades (e.g., B+ to A or A+), while 28 students went up by 7.5% or more test marks and that 18 students went down by two grades, while 15 students went down by 7.5% or more. Note the difference in changes for grades and marks because of the range of marks within each grade boundary.

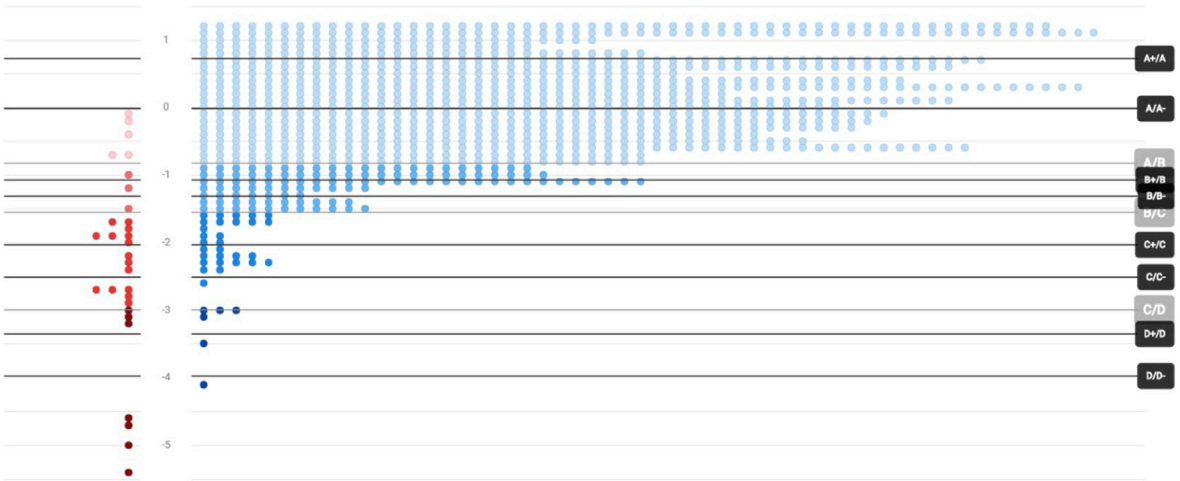
Thus, SmartStandardSet provides an industry-standard statistical tool for analyzing test questions and estimating student proficiency. The automated interface allows grade boundaries to be set and transforms scores into a format needed for institutional learning management systems. The system disregards the raw score, transforming it instead to reward proficiency with difficult items, and uses expert judgment to determine where grade boundaries should be set in light of the difficulty of content assessed. This standard setting approach avoids conventional approaches to adjusting scores for test quality (e.g., scaling scores to meet an expected grade distribution). Instead, it potentially focuses attention solely on the quality of performance (i.e., excellent, good, satisfactory, unsatisfactory) in response to item difficulty and content. Unsurprising given the potential impact on student scores and grades, SmartStandardSet displays the performance distribution of test takers before and after the calculations so that the impact of standard setting on students receiving all grades can be considered.

SmartStandardSet makes no assumption as to whether the test is administered digitally or on-paper. All that is required is that it is given data in a digital format for analysis purposes, which can be achieved by automated scanning of paper-based answer sheets. Finally, it should be noted that while letter grading is a common approach globally at the tertiary level and standard setting with respect to grade boundaries is core to the SmartStandardSet tool, the system could still be used in contexts where final scores are numeric. In such cases, the final standard setting step could be omitted and the computed 2 PL IRT scores could either be used directly as a measure of performance or transformed to a desired distribution or score range.

METHOD

Aim and Objectives

Nonetheless, despite the technical and efficiency advantages of SmartStandardSet it remains to be seen if end-users, especially



Grades

A+/A boundary	0.73	A/A- boundary	-0.01
B+/B boundary	-1.07333333333	B/B- boundary	-1.31666666667
C+/C boundary	-2.03666666667	C/C- boundary	-2.51333333333
D+/D boundary	-3.35	D/D- boundary	-3.97

FIGURE 4 | SmartStandardSet standard setting display: Augmented grade boundaries.

Summary

55 students up 2 or more grades		28 students up 7.5%+	
130 students 1 grade		206 students up 2.5 - 7.5%	
442 students the same grade		656 students ± 2.5%	
331 students 1 grade		71 students down 2.5 - 7.5%	
18 students 2 or more grades		15 students down 7.5%+	

Students

Student ID	Raw %	Raw grade	Analyzed %	Analyzed location	Recommended %	Recommended grade
975364506	96.7	A+	96.7	0.91	97.3	A+
257089802	70	B	70	-1.29	66.5	B
310529964	96.7	A+	96.7	0.73	95.5	A
247441715	96.7	A+	96.7	0.65	94.7	A
453266161	83.4	A-	83.4	-0.33	85	A-

FIGURE 5 | Standard setting display transformed grade display.

students, would evaluate the output and processes as accurate, useful, and fair. Consequently, this paper examines, in accord with core evaluation standards (Yarbrough et al., 2011), stakeholder perceptions concerning the accuracy of the statistical scoring methods, the utility of SmartStandardSet, the propriety of adjusting raw marks through IRT and standard-setting, and feasibility factors that might incentivize usage. Hence, this study pays explicit attention to maximizing end-user consequences (Sen, 2000).

Unfortunately, automated evaluation of applications is not yet feasible, and rubric guided evaluations (von Wangenheim et al., 2018) do not achieve high levels of consistency (da Silva Solecki et al., 2019). Of all user experience features, in a recent review of computational thinking assessments, only user interactivity was identified as something for which there was automation (Cutumisu et al., 2019). Hence, a non-experimental case study (Yin, 2006), using convenience samples of instructors and students within the researchers' own university was carried out. Potential participants were recruited from among course directors who were also course instructors known by the research team to use MCQ testing extensively. Six of the eight instructors approached chose to participate. These participants were then asked for permission to recruit students in their courses. Recruitment and procedures followed standard institutional review board requirements of informed consent and voluntary participation (HPEC 2018/021976).

Individual instructor interviews took place in spaces near their main department at mutually agreed times. Student focus groups likewise took place near the main department of the participants and at mutually agreed times. Discussion prompts for teachers and students focused on three major aspects, that is statistical analysis of test scores, standard setting protocols, and policy relationships (details in **Supplementary Appendices A,B**, respectively). These topics were selected to align with evaluation standards of accuracy, utility, and propriety. Refreshments were served and an honoraria of either $10 (instructors) or $25 (students) was paid [all in accordance with the Human Participants Ethics Committee (HPEC); #021976].

Participants

Per the World Bank (n.d.), New Zealand is a high-income country with a Human Capital Index = 0.77/1.00, which is equivalent to France and the United Kingdom. This study was carried out in a large ($N \approx 40{,}000$), publically-funded, research-intensive, comprehensive university, situated in the largest metropolitan region (approximately one-third of national population) of the country. Entry is selective in that students are required to have a minimum of 150 points from the best 80 credits earned in the New Zealand National Certificate of Educational Achievement Level 3. This contrasts to the minimum entry score of 120 points used at all other universities in the country.

Six individual instructors representing four different faculties were interviewed (**Table 1**). Sequential ID codes were assigned to participants (SIP1, SIP2, etc.) where the prefix "SI" indicates a staff interview and the suffix P1, P2, etc., indicates the order of the interviews. Six focus groups were conducted with 19 first- and second-year students from each of six faculties. Sequential ID codes were assigned to participants (P1, P2, etc.) and prefixed FG to indicate focus group participation.

Although we have provided some demographic information about participants we do not use that information to analyse results. The sample sizes of sub-groups are too small to generate stable estimates of difference and the type of data do not lend themselves easily to analysis of variance. This information instead illustrates the diversity of backgrounds and allows us the opportunity to establish variation in beliefs and attitudes across the sample.

Qualitative Data Analysis

Data-analysis followed Srivastava and Hopwood's (2009) qualitative iterative data-analysis model. The iterative process is a reflective systematic scheme with the aim of reaching understanding and meaning rather than just repeatedly coding text fragments until a pattern or a theme emerges. The insight is then associated with what the researcher knows, what he or she wants to know, and the logical relationship between that pre-existing knowledge and the material emerging from the data (Srivastava and Hopwood, 2009). **Figure 6** shows the iterative analytic procedure used in this study. The process involves first grouping responses based on whether the feedback was positive or negative. Subsequent re-examination was employed to search for possible new insights. The summarised data was then synthesised for an interpretation of user experience of SmartStandardSet's acceptance, feedback around implementation, policy, and impact to MCQ testing.

RESULTS

Data from across interviews and focus groups has been organised around core evaluation attributes of accuracy, utility, propriety, and feasibility. This allows the reader a more comprehensive understanding of how SmartStandardSet is perceived. **Figure 7** suggests in green that the accuracy and utility of SmartStandardSet were viewed positively. However, the orange box of standard setting and the grey field of policy both indicate caution concerning the propriety of implementing this technology.

Accuracy

Both instructors and students indicated high levels of acceptance for the credibility and trustworthiness of the statistical analyses conducted in SmartStandardSet. Instructors generally perceived that the scores created by SmartStandardSet were an accurate way of determining scores and understood how SmartStandardSet removed the guesswork in creating a credible statistically informed score.

> I'm a bit biased because I understand the statistics and the statistical analysis that can happen, so I find it quite credible because I know this sort of mathematical formulas. . . . I understand the rationale behind it

TABLE 1 | Participant characteristics.

Instructor interviewees

Code	Sex	Faculty
SIP1	Female	Medical and Health Sciences
sSIP2	Female	Education
SIP3	Male	Education
SIP4	Female	Education
SIP5	Female	Biology
SIP6	Female	Information Systems and Operations

Student Focus Groups

Group	Code	Sex	Year of Study	Department
A	FGP1	Male	1	Engineering
A	FGP2	Male	1	Engineering
A	FGP3	Male	2	Engineering
A	FGP4	Male	2	Engineering
A	FGP5	Female	2	Engineering
B	FGP6	Female	1	Biology
B	FGP7	Female	1	Biology
B	FGP8	Male	1	Biology
B	FGP9	Male	1	Biology
C	FGP10	Female	2	Education
C	FGP11	Female	2	Education
C	FGP12	Female	2	Education
C	FGP13	Female	2	Education
C	FGP14	Male	2	Education
D	FGP15	Male	1	Information Systems and Operations
D	FGP16	Female	1	Information Systems and Operations
E	FGP17	Female	2	Medical and Health Sciences
E	FGP18	Female	2	Medical and Health Sciences
F	FGP19	Female	2	Arts

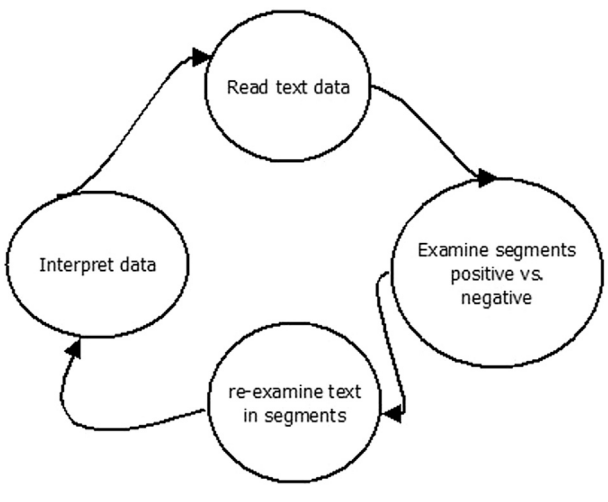


FIGURE 6 | Qualitative Data-Analysis Plan and the study's iterative analytic procedure.

and seems like a credible process. I have no problem with it. Yes, I think it's quite credible . . . I think it's definitely a credible process (SIP3).

It is very dependable.... it definitely better that what I was doing before. and I will definitely use it (SIP4).

It was noted by some that this application could do quickly what they would otherwise do manually. The efficient identification of deficient items within a test and the ability to delete those and automatically recalculate test scores was seen as a real benefit.

So, we're always trying to analyse our items after an MCQ test and we become very aware by looking at the student output, so its eyeballing. So, I think being able to pull those questions out and help improve question design is really important. . . . there's a huge advantage in removing deficient items from test. We do not have anything like this and would love to use something like this going forward (SIP5).

We've had that, we're a large course so I think we've had all of the above scenario, we've given a bonus mark to the class in that case so yeah, we've just admitted that this was poorly worded, we normally have all five people doing proof reads of the test but it happens and we just give them a bonus mark, so yeah it's good to have something like this (SIP6).

There's a huge advantage in removing statistically deficient items from test. We normally do this by analysing students' MCQ results and comparing it to the overall value. (SIP1).

Students likewise wanted instructors to use a system like SmartStandardSet to statistically evaluate MCQ items and

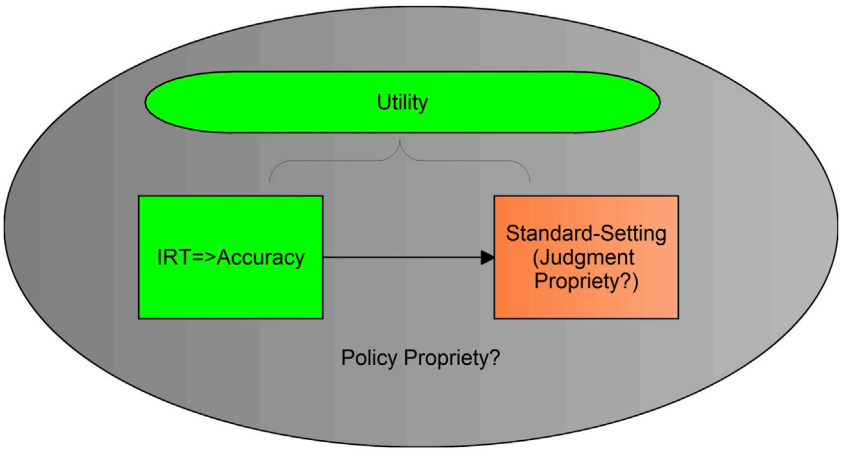


FIGURE 7 | Evaluation Overview. Note. Green indicates positive evaluation; orange indicates caution; grey indicates uncertainty.

remove deficient ones rather than following current MCQ test analysis protocols, in which all items are retained regardless of quality. Students considered it only fair that poor quality test items should be removed.

> The advantages of pulling items out because of the possibility of it not being taught correctly. Yes, especially if the teacher's question is flawed. Yeah, I will definitely accept this (FGP5).
>
> If the question is bad it should be removed. With this, teachers can know for sure which questions are bad . . . (FGP7).
>
> I think this is good and identifying and removing bad items is good. Yes, I would feel comfortable in removing bad items (FGP13).
>
> I think it's good because if you can take it [item no one gets right] out because if the people who are studying and can't get that question correct then it should be taken out because it is irrelevant to the test (FGP4).

Another feature that was evaluated positively by both instructors and students was that of giving more weight to difficult items.

> It could help to provide more challenging questions and . . . where students get that correct then they will be rewarded for that (SIP5).
>
> I think from a course perspective it's a really good idea, it's good to reward students who have been able to understand the extra bit (SIP6).
>
> Easy questions should be less points and difficult questions should have more points. I do find the system credible because I can see it right here (FGP9).

Students also suggested that weighting items might help to better differentiate students, a matter of some concern in highly competitive programs.

> It is really good too because it separates people from A to A plus and B to B plus (FGP3).
>
> What I like is that this brings in a differentiating factor. . . . But with this it does show differences between student abilities. So . . . A+ is really hard to get and you should be rewarded for earning that (FGP1).

The overwhelming feedback from participants was that the analysis and evaluation of MCQ items in SmartStandardSet is credible, dependable, and trustworthy and that all stakeholders can understand what is happening. Reduced guesswork in evaluating items was appreciated by academics in terms of workload and by students in terms of resistance of the system to potential subjectivity.

Feasibility

Feasibility refers to the effectiveness and efficiency of the innovation. After a brief demonstration and with no training, except for on-screen help functions, instructors deemed the system easy to use.

> SmartStandardSet is . . . easy to use . . . By the look of it yes.... it [is] definitely better that what I was doing before. And I will definitely use it (SIP4).
>
> The system feedback is good, and I would be pretty confident in using this system (SIP1).
>
> Yeah, I think we could do this it might be a good idea . . . I want to use this next year (SIP2).

They were pleased to see that through the use of standard data formats, SmartStandardSet achieved interactivity with other software systems (e.g., MCQ Results, Remark, etc.) used to process MCQ tests in their various faculties.

> I've used Crowdmark with the MCQ's, I've used Remark for quite a few semesters as well, so I think that would be quite easy to implement [SmartStandardSet], it would be no issue there (SIP6).

Given the design efforts and alpha-testing done prior to this study, we were satisfied that the software was perceived so positively.

Propriety

Three issues of propriety were identified: 1) ensuring SmartStandardSet protocols are consistent with current policy and regulations; 2) ensuring complete transparency about the protocols; and 3) the difference between standard-setting and scaling.

Policy and Regulations

A challenging feature of the SmartStandardSet protocols lay in whether they were permitted under current regulations. It was expected that explicit statements in the examination policy and procedures to allow 1) item weighting, 2) student score weighting, and 3) grading based on judgment of academics would be necessary. However, five of the six instructors did not see any challenges around using SmartStandardSet to remove deficient items to ensure test quality or around setting standards.

> No issues at all and once again, a great synergy with other MCQ software analysis programme I currently use now (SIP1).
>
> I don't see a problem with that. To be honest no in terms of policy if that fits with the ethos of what the university requires for its assessment principles, then it's all good with me. It should, and I believe it does, assess true learning in a fair way (SIP2).
>
> No issues at all. I don't think this is a policy issue because you are working on marks earned by students and creating difficult and challenging questions and testing student's knowledge is a part of it right? So, no issues at all (SIP4).
>
> If you're focusing on the standards then [you're] getting a true picture where the students are meeting that standard, not just your simple statement of a standard but within the detail that they're addressing that which comes from recognising that not all questions are equal (SIP5).
>
> As far as I'm concerned, well, there's no regulation as such . . . I've used Crowdmark with MCQ's, I've used Remark for quite a few semesters as well, so I think that would be quite easy to implement, it would be no issue there (SIP6).

Unsurprisingly, some students were reluctant to let instructors set grade standards other than implement the conventional mapping of percentage scores to letter grades.

> I fully understand what you are saying; but no (FGP15).
>
> Yup, once again I fully understand the reasoning behind it; but I don't know (FGP5).
>
> But, not a lot of students will like it. Just saying (FGP16).
>
> I totally understand how this works and I know people will not like it (FGP19).

It seems these students, despite the accuracy and utility features of the software, were still concerned about the legitimacy of making these changes. Resistance to assessment novelty is, of course, well-established in the literature (Struyven and Devesa, 2016).

Although there was hesitancy concerning this change, some students indicated that well-informed students (e.g., those who know about statistics) might actually accept this innovation.

> I think university students will accept this and especially if the teachers tell students before the exam then they will accept it. Especially, if they know something about statistics (FGP16).

Indeed, a few students suggested that this kind of judgment-based adjustment of exam and test scores is already part of the landscape.

> Well, for med-sci [medical science] this is actually sort of happening (FGP17).
>
> With this . . . some questions are more difficult or easy and the harder get more points. Some of my lecturers do that already. So, if that's what you are talking about, then yeah it is in accordance with the university (FGP5).

Not unsurprising, if adopted, one student pointed out academics have to be adequately trained to use and understand the logic of standard-setting.

> Policy-wise and if they are implementing this, they should make sure that everyone is using this and are trained accordingly (FGP8).

Concern that academics may not be valid or accurate in their standard setting was identified. Without transparency about how cut-scores were set and surety that all instructors were using the same standards, there would be some potential for mistrust.

> I'm not sure about this. I mean fundamentally I understand how it works . . ..but if teachers have to set it? I am not sure. Maybe, have them do it in front of class, but if they don't' tell us then they will need to have better explanation (FGP18).
>
> Implement it campus-wide and make sure teachers are trained correctly. It would be nice of the actual teacher with the people who are in-charge of training them to be there also to show us all how it works and how you guys are implementing it (FGP16).

Transparency

Two instructors pointed out that students have a right to know how their tests are scored and analysed already, so informing them of this variation would be essential.

> I've done this. So, we've got rules about dissemination and communicating those results and how did I get those results. So, whatever you do as long as you declare it to students and there's no objection to it, I don't think there are any issues to be honest (SIP2).
>
> You have to be transparent to students about what's happening . . . I'm always honest with the students and say it's difficult to write good questions and sometimes [they are] literally are sh*t and you can't answer them. . . . I just explain to them that the system will pull out the ones that are not reliable and not fair in that way but also it will look at the ones that are more challenging and it will adjust the grade based on that (SIP5).

Students themselves suggested that, with a high-level of transparency and feedback about which items were removed and which were hard or easy on a test, there might be more acceptance of the protocol.

> If the question is bad, it should be removed. With this, teachers can know for sure which questions are bad and should be talked about its class. A good explanation why it's bad and why it should be removed (FGP7).
>
> I do want lecturers to show me how they are analysing this. Some explanation of which questions are good and bad and the questions removed (FGP10).
>
> There are several different advantages for removing bad questions–I mean it's good that teachers can look at the questions and can know which ones are good and which ones are bad. As long as they are transparent . . . (FGP12).
>
> Also, I would like to get overall feedback on questions . . . like, how many students got specific questions correct and incorrect. So, each question I would like the lecturer to breakdown the overall percentages (FGP18).

Students suggested that knowing how the test was marked, which items were deleted for which reasons, and why items were easy or hard would assist them with their learning, as well as making the practice acceptable.

However, in the time available, we were unable to ascertain how students might view two different kinds of bad questions (i.e., those all got right vs. those none got right). We might expect students to endorse removing items that were too hard for the whole test cohort as that would reduce the divisor in creating a percentage score. However, the legitimacy of removing an item that all got right may seem to be unfairly punishing performance, instead of improving the quality of a test through deletion of non-discriminating items. This is clearly a matter meriting further investigation.

Scaling vs. Standard-Setting

Another challenge lay in whether the standard-setting protocol was different to the traditional scaling of marks to achieve certain grade distributions. SmartStandardSet requires instructors to judge grade standards according to known properties of their test, while having access to the student distribution of proficiency. Three instructors argued that standard-setting based on item quality or content relative to university grade standards was different to scaling.

> Well, scaling to me is artificial because it says ok, you got 55 percent so you're in the top 5% then you should get 85%; artificial. But for someone, because of the difficulty of a question having more marks, that's not artificial. That's real. You know they've earned them; that it's not artificial. So, I don't have an issue in terms of scaling concept. It's a reflection of their true learning. It's a sophisticated and complex concept that has required deep understanding and synthesis of thought (SIP2).
>
> I think it's acceptable, because at the end of the day we've got learning outcomes and grading scenes and criteria of what constitutes A, B and C, so I think having a standard makes it quite easy to justify why you belong in the A range and not the B and not the C and vice versa. So, it's probably a combination of both rather than one or the other . . . I don't think, as I said before, you're not taking away marks or awarding marks for certain people, you are adjusting for the whole class and because it's a whole class approach, I don't see anything wrong with it. . . . I have no problem with it (SIP3).
>
> I think a lot of students get very good at evaluating the amount of preparation they need to get an A+ and that's all they do, and you can understand that in terms of the demands of their time. However, if the way the results are analysed and adjusted afterwards requires them to have to do those harder questions to get the higher grades . . .. then they will put the work on and it is a more reliable A- [or] A+ and they don't want to know that they got an easy A- [or] A+. They want to know that it is, genuinely very reflective of a high standard (SIP5).
>
> However, not all instructors were convinced: It's pretty much scaling and it does not adhere to the university's policy (SIP1).

Thus, considerable effort would need to be made to clarify the distinction between scaling and standard-setting, instantiate the protocol as a legitimate option within the assessment policy, and persuade students that the grades arrived at in this fashion are defensible. Nonetheless, any significant change in assessment practice is a major challenge.

Utility

Utility of an innovation depends on the value intended users perceive in making use of it. The discussions touched on potential benefits for 1) the quality of tests written by instructors, 2) the quality of in-class questioning, 3) how students approach test-taking, and 4) how grades are determined.

Impact on Test Writing

The real usefulness of SmartStandardSet was seen in its ability to give test item writers feedback as to whether their items were

positively discriminating and as hard or easy as they had predicted in the writing phase.

> Absolutely, very useful . . . It gives me feedback about how I've written the item, it gives me feedback about how students might have thought about the item and why I didn't, you know it gives you that student perspective which you don't usually get if there's no, none of that kind of process (SIP5).

One instructor noted that realizing a test has mostly quite easy items, might lead to greater efforts to ensure more challenging items are included in future tests.

> I think the other thing that will be highly valuable, a bit subversive [is] it will make the teaching staff pay attention to the questions they're putting out. I think it's very easy to just pick a few, tweak them around the edges and put them into the [test] script and so forth . . . so I think it could also mean that the standard of questions improves in the script with this opportunity to have the more difficult ones and the students rewarded for that (SIP5).

The presumption of IRT is that if students get a lot of easy items correct, then despite this high percentage of correct responses, their ability is not necessarily advanced or highly proficient. Thus, SmartStandardSet has the potential to influence instructors to write harder and more challenging MCQs.

Impact on Teaching Practice

A fairly standard practice post-examination is to have a review of items. While this is potentially a transparency requirement, a student noticed that SmartStandardSet would make this much easier.

> My teachers normally just use and show class average and show which questions are difficult or easy based on how many got the question incorrect vs. correct. This gives a statistical tool for teachers to use and this makes so much more sense (FGP10).

At least one instructor explicitly indicated that because SmartStandardSet gives higher scores to students who answer harder questions, there might be a spin-off effect on their teaching. If more challenging questions result in higher marks, then instructors would have to model such questioning within their own teaching. In this case, the utility of SmartStandardSet is not just for future testing, but also for future teaching.

> It does reward the people who can think who have developed a critical thinking, who have really looked into a deep understanding of the thing . . .. I might be motivated to teach more carefully. I mean I do try to ensure and understand the concepts of the course. We do that hopefully, but I might, it might make me reflect on my teaching too a little bit and hopefully it will make them reflect on how they are learning (SIP2).

Hence, SmartStandardSet, by making easily accessible item discrimination and difficulty indices, has potential use for improving the quality of test item writing and even possibly course instruction. Knowing how one's items perform acts as a prompt for writing challenging and difficult questions that are more likely to ensure evidence of excellent performance.

Impact on Test-Taking

A surprising result of the weighted score method was the positive view students took of this in terms of their test-taking strategies. They indicated greater willingness to spend more time on difficult questions so that they could be rewarded for getting those items right. This is instead of the current strategy of skipping hard items because there is little risk in doing so.

> This will highly affect behaviour in the exam itself... I think. Normally, students will skip difficult questions . . . they are all worth one-point anyways. But if students know that certain questions or difficult questions weigh more then you will spend more time with that question (FGP5).
>
> Honestly, students will understand how this works. Some will probably still do the easy questions first, but they will for sure go back and try to answer the difficult questions because they know it's worth more (FGP6).
>
> Yes, it's a cost sometimes when you spend more time on a question and you will probably get it right, but you have less time for the other questions. With this, students are rewarded. You are rewarded for spending time and effort (FGP2).
>
> Yeah, I will probably spend more time with difficult questions because I get paid off at the end and especially, if I get it correct. I mean, I should get the easy questions correct right so I am not worried about that (FGP3).

Because SmartStandardSet gives more weight to harder items which should require deeper more complex cognitive processing, instructors seemed to think that this would potentially change how students approach learning and testing. Instead of rushing through items in the hope of getting as many correct answers as possible, students ought to concentrate on items and tasks that are more difficult. The score system is designed to reward students for mastering deeper, more complex learning.

> Students might be motivated to read carefully and deeply . . .. And then of course in tutorials I give them exemplars and trial questions I could point out what might constitute a difficult question as opposed to an easy question (SIP1).
>
> In principle, if students knew the parameters of the thing and how it was going to be marked, you know they

might be motivated to read carefully and deeply. . . . I wonder if students knew that, if they would be encouraged to persevere with a deeper meaning of the course material. . . . It's a reflection of their true learning. It's a sophisticated and complex concept that has required deep understanding and synthesis of thought. John Hattie says that a good MCQ should test just as well as any other means (SIP2).

I can easily see reasoning behind difficult and weighted items and the discrimination between high and low ability students. If I explain this to my students, then they will probably focus on every question and take their time (SIP4).

So, I think it could actually be a cool way to start to get them to prepare differently for multiple choice, because there's a huge amount of criticism for using multi choice and it's exactly that, its surface learning but this has the potential (SIP5).

One student explicitly made the same point.

Yeah, I will probably spend more time with difficult questions because I get paid off at the end and especially, if I get it correct. I mean, I should get the easy questions correct right so I am not worried about that (FGP3).

Impact on Grading

Students indicated that they valued the approach of weighting items before getting a grade. From their perspective, greater rewards for students who have mastered the hardest material means that there is greater discrimination and differentiation when course grades are issued.

It is really good too because it separates people from A to A plus and B to B plus (FGP3).

What I like is that this brings in a differentiating factor. Students who are getting a D is the same as someone who is getting a C. But with this it does show differences between student abilities. So, A and A+ shows differences. A+ is really hard to get and yeah you should be rewarded for earning that (FGP1).

Surprisingly, some of the students were not very sympathetic to weaker students whose scores would go down if they could only answer the easy items.

Some students will not like it but that's tough right? Easy questions should be less points and difficult questions should have more points. I do find the system credible because I can see it right here (FGP9).

Yeah, that's the way it has to be. I mean if this goes up for them and this goes down for them that's pretty much how it goes. Most likely they didn't study for that question or they are incapable (FGP10).

I am not sympathetic to people who don't study . . . Including myself. If I don't study and don't get the grade, then I deserve that (FGP14).

If I am this student whose grade was dropped, I will be sad, but it does tell me something about how I am studying and maybe improve the way I am studying. Or, improve my understanding (FGP11).

If implemented, it would appear that SmartStandardSet has potential to not just improve test writing but also to have a positive effect on teaching, student test-taking strategies, and ultimately create grades that meaningfully identify quality rather than quantity.

DISCUSSION

This case study used potential end-user perceptions to evaluate a new software tool that uses an IRT test-scoring algorithm combined with instructor judgment to set grade standards. The goal was to determine if 1) university lecturers and 2) university students would accept the new software system as a valid, accurate, useful, and ethical alternative to the conventional approach to test scoring and converting to letter grades. The evaluation took place in a research-intensive university in which MCQ tests are commonly used, especially in science, technology, engineering, and medical subjects, to make high-stakes decisions about student learning (i.e., scholarships, prizes, entry to graduate school, etc.).

The data (i.e., instructor interviews and student focus groups) support the conclusion that the software met participant expectations for feasibility, accuracy, and even utility. The IRT score adjustment for item difficulty was understood and accepted; the software interface and interoperability with institutional examination data files meant that the system was seen as being very useful; the mechanisms for setting scores and checking the impact of such decisions was seen as desirable.

However, issues were raised concerning the propriety of introducing a new protocol for determining grades, especially without formal policy approval from the institution. Nonetheless, students surprised us by indicating that they favoured the idea of being awarded higher grades for correctly answering harder questions. Clearly, a number of regulatory approvals or clarifications need to be put in place before wide-spread deployment of SmartStandardSet can be contemplated. Nevertheless, there is sufficient merit in the responses of these instructors and students to support further piloting of the system, perhaps initially around mid-term tests or weekly quizzes that have lower summative weight than final exams.

A clear limitation of this study is that it took place in one university and with a small number of informants. Generalizability to other institutions and acceptance across all disciplines that use MCQ testing needs to be established with further use and evaluation. The student sample distortion in terms of faculty (i.e., 50% in either engineering or education) may also limit applicability of results for students in humanities or arts. Similarly, the convenience process with instructors produced a skewed sample (i.e., 50% in education). Unfortunately, data on

student overall academic ability was not available, so we cannot eliminate the possibility that the responses observed here arise from highly successful students. This leaves open the possibility that less proficient or struggling students may have quite different reactions to this mechanism. Perhaps, more importantly, the staff and students were presented with a functioning software but not an operational testing situation. They were presented with real data, but not from a test that the instructors had administered or that the students had experienced. Future research needs to establish the impact of the IRT and standard setting decisions on a real test that has real consequences for instructors and learners.

Thus, this report should be seen as an exploratory, pilot study calling for robust experimental studies in which both students and instructors are assigned to conditions in which tests and grades are manipulated with SmartStandardSet versus conventional mechanisms. Open access to the SmartStandardSet application, which is being implemented, should generate interesting data about the quality of higher education MCQ tests, as well as insights into the acceptability of statistically weighted scores and instructor judged standards.

It is interesting to consider the potential of the IRT approach to scoring items on how instructors teach and write test items and on how students take tests and prepare for them. By ensuring that the content and skills associated with excellence are embedded in tests, our informants suggest that learners would concentrate on the more challenging tasks and will develop a strategy of investing resources in the hard test items. This potentially moves learners from the habits of surface learning (i.e., reproducing taught material) to a more effective deep learning strategy (i.e., transforming taught material into new knowledge). Should this actually arise, then grades would more explicitly reflect qualities of excellence instead of quantities. While this is clearly a desirable outcome, it is dependent on the extent to which students can accurately identify the difficulty of the items on a test. Traditionally, test construction practices tend to put the easiest items at the beginning of a test and from that pattern, students may be able to discern the hardest items. Furthermore, it would be possible for the teacher to assist students by indicating in some way which questions are easy and which are difficult. However, we argue that such analysis is best left to the student. The ability to make accurate evaluative judgments is an important metacognitive skill (Schunk, 2008; Tai et al., 2018; Prather et al., 2020) and there is evidence that encouraging students to reflect on the difficulty of questions provides measurable benefits. For example, Denny et al. (2010) found that when students are prompted to regularly practice assessing their confidence in a multiple-choice question answer–a proxy for question difficulty–they outperform students who do not receive such explicit prompting. Weighting question scores by their difficulty, as supported by SmartStandardSet, provides an implicit incentive for students to work on developing their self-assessment skills.

An interesting study, then, would be to examine if students can, under test conditions, reliably predict which items are more difficult and whether spending more effort on those items has a substantial impact on their score and item psychometric properties. While it is possible that item parameters could shift if all students were to succeed on "harder" items, it seems likely that only more proficient students would have success in this way. Thus, a useful study would be to examine the accuracy of student detection of item difficulty and the consequence of greater effort at the item level relative to a student's general ability.

While innovations in assessment are generally resisted by students (Struyven and Devesa, 2016), this study suggests that the novelty introduced in SmartStandardSet appeals to students' sense of fairness. To be rewarded for knowing the harder material, a matter disguised by CTT scoring methods, and to be more selectively awarded higher grades appealed to these students. Nonetheless, their endorsement of IRT scoring on grounds of fairness does incriminate higher education's conventional CTT approaches to evaluating students. Further work with broader samples of students would test whether less proficient or marginal students have the same attitude towards this type of change.

The interesting challenge for higher education that this study addresses relates to ensuring that assessments are high quality. While it is fashionable to decry the validity of MCQ testing, it is clear that their place, in courses with large classes and in technical disciplines with large bodies of compulsory knowledge to be acquired, is assured. Nonetheless, unlike the extensive use of IRT test scoring in admissions testing, international large-scale testing, and standardized tests in K-12 systems, higher education institutions lag severely behind in how they analyse MCQ tests (Brown and Abdulnabi, 2017). It is more than ironic that, while IRT has been developed, taught, and refined in universities, they fail to make use of such tools in their own practice. The difficulties of implementation have been overcome thanks to the use of open-source software, the use of a well-established report template, and the willingness of instructors to exercise their professional judgment about quality.

FUTURE WORK

In this exploratory work, we present rich qualitative descriptions of how instructors and students viewed the accuracy, feasibility and utility of SmartStandardSet. We considered it prudent to explore these views in a pilot study, using authentic but historical course data, before application in a real course where actual student grades are at stake. While we believe this cautious approach was justified, pilot studies tend to be small in scale and represent only an initial step in a broader evaluation. Although our initial results are promising, future work is needed to explore how these findings generalize to other institutional contexts and, importantly, how perceptions vary when the impact to student grades is real.

We interviewed six instructors and ran six focus groups with 19 students all from the same institution, albeit recruited from six different faculties. One avenue for future work is to explore instructor and student perceptions at a much larger scale–recruiting participants from multiple institutions and collecting quantitative responses to questionnaires to measure objectively how perceptions vary across discipline area and student ability. A further advantage is that questionnaire data can be collected anonymously, potentially yielding more truthful responses. For example in this pilot study, the interviewees knew the broader research team. Although they did not know the

individual researcher conducting the interviews, the face-to-face nature of the interviews and the lack of anonymity of their responses may have limited feedback of a critical nature. Another stakeholder group not yet canvassed are university administrators and academic leaders. While the funding body representatives were positively impressed with SmartStandardSet, it remains to be seen how well this system would be accepted for operational purposes. Policy makers would have to be persuaded that the system achieves desirable goals with external stakeholders (i.e., employers, parents) before being implemented. Hence, it is highly likely that several pilot studies, in which operational test or examination scores can be evaluated conventionally and with SmartStandardSet for acceptability, would be required.

One aspect of IRT that was not built into SmartStandardSet was model fit information. Goodness of fit tests (GoF) provide information about whether the test is unidimensional and which items fit well with that assumption. However, GoF tests may not actually provide much useful information for test users. Maydeu-Olivares (2013) makes clear that "In practice, it is likely that the fitted model [would] be rejected using an overall GOF statistic. Simply, it is not easy to find the data-generating model" (p. 72). By definition, all models are sufficiently discrepant from the data because of simplification processes in creating a model (e.g., 2 PL models explain performance and ability with just two parameters). Thus, it is quite easy not to find correspondence between the model and the data. Other research has suggested that using more sophisticated 2 PL or 3 PL models when the simpler 1 PL model is correct are unlikely to create more misfit (McKinley and Mills 1985; Jiao and Lau, 2003). This corresponds with Brown and Abdulnabi (2017) who reported that in the MCQ tests they analysed, the 2 PL model had the best fit. From this, we conclude that defaulting to 2 PL is unlikely to create much misfit of the model to the data, while providing novel and useful diagnostic information about the items. Hence, providing additional information at this stage of development may be premature. Nevertheless, provided development and training budgets permit such information could be provided, leading to interesting research questions as to how users understand model fit information.

To facilitate future work, one of the contributions we make in this study is provision of the SmartStandardSet tool as open-source software.[1] We encourage other researchers to build on this software and customize it for their own use and evaluation purposes. For example, grade boundaries (as shown in **Figure 4**) can be modified to suit any local institutional context. Goodness of fit measures for the IRT model could be included in the interface and assessed from a usability perspective.

A final avenue for future work, necessary prior to any widespread adoption of SmartStandardSet, would be to use the tool in a real course where student grades are impacted. Students, quite naturally, can be resistant to assessment changes that will affect their course grades. Despite the positive student feedback we observed in this pilot study, particularly around a desire to be rewarded for answering difficult questions correctly, it is important to further test these beliefs in an authentic classroom scenario.

[1]SmartStandardSet Github https://github.com/genericity/test-analysis

CONCLUSION

SmartStandardSet overcomes some of the barriers to using IRT scoring in higher education MCQ testing by providing an easy-to-use, open-source tool for instructors. Generally, its mechanisms were positively evaluated for accuracy, feasibility, and utility. Understandable challenges need to be addressed in order to allay concerns around propriety. Further investigation is required; nevertheless, the software seems to be ready for field trials where instructors use it for evaluating test quality and experiment with the possibility of setting grade standards according to their own judgement.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The University of Auckland Human Participants Ethics Committee #021976. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

GB and PD conceived of software; EL wrote and tested the software. GB and PD conceived evaluation study. DS conducted field work, transcribed and performed first analysis of data, and wrote technical report. GB and PD wrote, edited, and finalised manuscript.

FUNDING

This work was supported by The University of Auckland Learning Enhancement Grant system (grant number 70630).

ACKNOWLEDGMENTS

The SmartStandardSet software was developed by EL, then a final year Software Engineering student. DS is thanked for his extensive work on the interview and focus group data collection and analysis. Appreciation to all participants for their input.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2021.735088/full#supplementary-material

REFERENCES

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behav. Hum. Decis. Process.* 50 (2), 179–211. doi:10.1016/0749-5978(91)90020-T

Angoff, W. H. (1974). *Criterion-referencing, norm-referencing, and the SAT (Research Memorandum RM-74-1)*. Princeton, NJ: Educational Testing Service.

Baldwin, P., Margolis, M. J., Clauser, B. E., Mee, J., and Winward, M. (2020). The Choice of Response Probability in Bookmark Standard Setting: An Experimental Study. *Educ. Meas. Issues Pract.* 39 (1), 37–44. doi:10.1111/emip.12230

Brown, G. T. L., and Abdulnabi, H. H. A. (2017). Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. *Front. Educ.* 2 (24). doi:10.3389/feduc.2017.00024

Butler, A. C. (2018). Multiple-Choice Testing in Education: Are the Best Practices for Assessment Also Good for Learning?. *J. Appl. Res. Mem. Cogn.* 7 (3), 323–331. doi:10.1016/j.jarmac.2018.07.002

Capan Melser, M., Steiner-Hofbauer, V., Lilaj, B., Agis, H., Knaus, A., and Holzinger, A. (2020). Knowledge, application and how about competence? Qualitative assessment of multiple-choice questions for dental students. *Med. Educ. Online* 25 (1), 1714199–1714208. doi:10.1080/10872981.2020.1714199

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for theREnvironment. *J. Stat. Soft.* 48 (6), 1–29. doi:10.18637/jss.v048.i06

Cutumisu, M., Adams, C., and Lu, C. (2019). A scoping review of empirical research on recent computational thinking assessments. *J. Sci. Educ. Technol.* 28, 651–676. doi:10.1007/s10956-019-09799-3

da Silva Solecki, I., da Cruz Alves, N., Porto, J. V. A., von Wangenheim, C. G., Justen, K. A., Borgatto, A. F., et al. (2019). *Codemaster UI Design - App Inventor: A rubric for the assessment of the interface design of android apps developed with app inventor*. Vitoria ES, Brazil: Paper presented at the IHC '19.

Denny, P., Luxton-Reilly, A., Hamer, J., Dahlstrom, D. B., and Purchase, H. C. (2010). "Self-predicted and actual performance in an introductory programming course," in *Proceedings of the Fifteenth Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE '10)* (New York, NY, USA: Association for Computing Machinery), 118–122. doi:10.1145/1822090.1822124

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Adv. Health Sci. Educ. Theor. Pract* 10 (2), 133–143. doi:10.1007/s10459-004-4019-5

Embretson, S. E., and Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: LEA.

G. J. Cizek (2001). *Setting performance standards: Concepts, methods, and perspectives* (Mahwah, NJ: Lawrence Erlbaum Associates).

Halpern, D. F., and Butler, H. A. (2013). "Assessment in higher education: Admissions and outcomes," in *APA handbook of testing and assessment in psychology*. Editor K. F. Geisinger (Washington, DC: American Psychological Association), Vol. 3, 319–336. doi:10.1037/14049-015

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hattie, J. A., Brown, G. T. L., Ward, L., Irving, S. E., and Keegan, P. J. (2006). Formative evaluation of an educational assessment technology innovation: Developers' insights into assessment tools for teaching and learning (asTTle). *J. MultiDisciplinary Eval* 5 (3), 1–54.

Hattie, J. A., and Purdie, N. (1998). "The SOLO model: Addressing fundamental measurement issues," in *Teaching and Learning in Higher Education*. Editors B. Dart and G. Boulton-Lewis (Melbourne, Aus: ACER), 145–176.

Herzog, J. B., Herzog, P. S., Talaga, P., Stanley, C. M., and Ricco, G. (2019). Providing Insight into the Relationship between Constructed Response Questions and Multiple Choice Questions in Introduction to Computer Programming Courses. *IEEE Front. Edu. Conf. (Fie)*, 1–5. Covington, KY, USA. doi:10.1109/FIE43999.2019.9028548

ICSEE (2015). Writing guidelines for classroom assessment. *Appl. Meas. Edu.* 15 (3), 309–334. doi:10.1207/S15324818AME1503_5

Jiao, H., and Lau, A. C. (2003). *The effects of model misfit in computerized classification test*. Chicago, IL: Paper presented at the annual meeting of the National Council of Educational Measurement. https://bit.ly/3uEwiDI.

Joshi, P. K., Jian, Y., Khunyakari, R., and Basu, S. (2019). *A novel alternative to analysing multiple choice questions via discrimination index*. arXiv Physics arXiv:1906.07941.

Katz, I. R., and Gorin, J. S. (2016). "Computerising assessment: Impacts on education stakeholders," in *Handbook of human and social conditions in assessment*. Editors G. T. L. Brown and L. R. Harris (New York: Routledge), 472–489.

Kolen, M. J. (2006). "Scaling and norming," in *Educational measurement*. Editor R. L. Brennan. 4th ed. (Westport, CT: Praeger Publishers), 155–186.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Meas. Interdiscip. Res. Perspect.* 11 (3), 71–101. doi:10.1080/15366367.2013.831680

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Med. Teach.* 26 (8), 709–712. doi:10.1080/01421590400013495

McKinley, R. L., and Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Appl. Psychol. Meas.* 9 (1), 49–57. doi:10.1177/014662168500900105

Mitzel, H. C., Lewis, D. M., Patz, R. J., and Green, D. R. (2001). "The bookmark procedure: Psychological perspectives," in *Setting performance standards: Concepts, methods, and perspectives*. Editor G. J. Cizek (Mahwah, NJ: Lawrence Erlbaum Associates), 249–281.

Prather, J., Becker, B. A., Craig, M., Denny, P., Loksa, D., and Margulieux, L. (2020). "What Do We Think We Think We Are Doing?," in *Proceedings of the 2020 ACM Conference on International Computing Education Research (ICER '20)* (New York, NY, USA: Association for Computing Machinery), 2–13. doi:10.1145/3372782.3406263

S. Blömeke and J.-E. Gustafsson (2017). *Standard Setting in Education: The Nordic Countries in an International Perspective* (Cham, Switzerland: Springer).

Schauber, S. K., and Hecht, M. (2020). How sure can we be that a student really failed? on the measurement precision of individual pass-fail decisions from the perspective of Item Response Theory. *Med. Teach.* 42 (12), 1–11. doi:10.1080/0142159X.2020.1811844

Scherer, R., Siddiq, F., and Tondeur, J. (2019). The technology acceptance model (TAM): A meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education. *Comput. Edu.* 128, 13–35. doi:10.1016/j.compedu.2018.09.009

Schunk, D. H. (2008). Metacognition, Self-regulation, and Self-regulated Learning: Research Recommendations. *Educ. Psychol. Rev.* 20, 463–467. doi:10.1007/s10648-008-9086-3

Sen, A. (2000). Consequential evaluation and practical reason. *J. Philos.* 97 (9), 477–502. doi:10.2307/2678488

Srivastava, P., and Hopwood, N. (2009). A practical iterative framework for qualitative data analysis. *Int. J. Qual. Methods* 8 (1), 76–84. doi:10.1177/160940690900800107

Struyven, K., and Devesa, J. (2016). "Students' perceptions of novel forms of assessment," in *Handbook of Human and Social Conditions in Assessment*. Editors G. T. L. Brown and L. R. Harris (New York: Routledge), 129–144.

Tai, J., Ajjawi, R., Boud, D., Dawson, P., and Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *High Educ.* 76 (3), 467–481. doi:10.1007/s10734-017-0220-3

Teo, T. (2010). Examining the influence of subjective norm and facilitating conditions on the intention to use technology among pre-service teachers: A structural equation modeling of an extended technology acceptance model. *Asia Pac. Educ. Rev.* 11 (2), 253–262. doi:10.1007/s12564-009-9066-4

Ventouras, E., Triantis, D., Tsiakas, P., and Stergiopoulos, C. (2010). Comparison of examination methods based on multiple-choice questions and constructed-response questions using personal computers. *Comput. Edu.* 54 (2), 455–461. doi:10.1016/j.compedu.2009.08.028

Wangenheim, C. G. v., Hauck, J. C. R., Demetrio, M. F., Pelle, R., Cruz Alves, N. d., Barbosa, H., et al. (2018). CodeMaster - Automatic Assessment and Grading of

App Inventor and Snap! Programs. *Inform. Edu.* 17 (1), 117–150. doi:10.15388/infedu.2018.08

Wright, B. D., and Bell, S. R. (1984). Item Banks: what, Why, How. *J. Educ. Meas.* 21 (4), 331–345. doi:10.1111/j.1745-3984.1984.tb01038.x

Wright, B. D., and Stone, M. H. (1979). *Best Test Design*. Chicago, IL: MESA press.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., and Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users*. 3rd ed. Thousand Oaks, CA: Sage.

Yin, R. K. (2006). "Case study methods," in *Handbook of complementary methods in education research*. Editors J. Green, G. Camilli, and P. Elmore (Mahwah, New Jersey: Lawrence Erlbaum Associates), 111–122.

Yousefi Afrashteh, M. (2021). Comparison of the validity of bookmark and Angoff standard setting methods in medical performance tests. *BMC Med. Educ.* 21 (1), 1–8. doi:10.1186/s12909-020-02436-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.