Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations?

Jonathan D. Kibble¹² and Teresa Johnson²

¹Health Sciences Centre, Memorial University of Newfoundland, St John’s, Newfoundland, Canada; and ²College of Medicine, University of Central Florida, Orlando, Florida

Submitted 6 July 2011; accepted in final form 19 August 2011

Kibble JD, Johnson T. Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? Adv Physiol Educ 35: 396–401, 2011; doi:10.1152/advan.00062.2011.—The purpose of this study was to evaluate whether multiple-choice item difficulty could be predicted either by a subjective judgment by the question author or by applying a learning taxonomy to the items. Eight physiology faculty members teaching an upper-level undergraduate human physiology course consented to participate in the study. The faculty members annotated questions before exams with the descriptors “easy,” “moderate,” or “hard” and classified them according to whether they tested knowledge, comprehension, or application. Overall analysis showed a statistically significant, but relatively low, correlation between the intended item difficulty and actual student scores (r = −0.19, P < 0.01), indicating that, as intended item difficulty increased, the resulting student scores on items tended to decrease. Although this expected inverse relationship was detected, faculty members were correct only 48% of the time when estimating difficulty. There was also significant individual variation among faculty members in the ability to predict item difficulty (x² = 16.84, P = 0.02). With regard to the cognitive level of items, no significant correlation was found between the item cognitive level and either actual student scores (r = −0.09, P = 0.14) or item discrimination (r = 0.05, P = 0.42). Despite the inability of faculty members to accurately predict item difficulty, the examinations were of high quality, as evidenced by reliability coefficients (Cronbach’s α) of 0.70–0.92, the rejection of only 4 of 300 items in the postexamination review, and a mean item discrimination (point biserial) of 0.37. In conclusion, the effort of assigning annotations describing intended difficulty and cognitive levels to multiple-choice items is of doubtful value in terms of controlling examination difficulty. However, we also report that the process of annotating questions may enhance examination validity and can reveal aspects of the hidden curriculum.

Bloom’s taxonomy; assessment; evaluation; multiple-choice questions; standard setting; physiology education; medical education; hidden curriculum

The need to set standards of educational achievement is a challenge common to all educational levels. Irrespective of what grading system is used, there is usually the need to identify single score points that differentiate between meaningful levels of competence or achievement. This is particularly true for tests such as multiple-choice examinations, where the difference between getting a single question right or wrong may have no real meaning! In a perfect world, we would always be able to pretest examination items to determine their difficulty and statistical characteristics before using them in high-stakes summative assessment. In the real world, however, this is not always feasible, and, therefore, the ability to predict these variables before setting tests is an attractive notion. There is some evidence that this can be done for math questions (9) but is very difficult to achieve, for example, in tests of writing ability (1). Under circumstances where teachers have the freedom to define a passing score, there are several elegant standard-setting methods available (6, 12). However, in many higher education institutions, including our own, we do not have the freedom to vary the passing score; rather, there is an arbitrary percentage that differentiates one grade from another. The practical consequence for faculty members is the need to craft examinations that fairly measure student competency and, at the same time, produce an acceptable grade distribution. Success in this endeavor is clearly related to the ability of faculty members to develop test items of appropriate difficulty.

The goal of this study was to assess how useful two multiple-choice question classification schemes would be in helping us to predict the difficulty of test items. In our undergraduate human physiology course, we hypothesized that faculty members would be able to make subjective predictions about the difficulty of their test items on a three-point scale ("easy," "moderate," and "hard"). Since it is common practice to apply learning taxonomies to classify multiple-choice questions (5, 8), we further hypothesized that the cognitive level of items could be used to predict item difficulty.

METHODS

The Faculty of Medicine of The Memorial University of Newfoundland offers a two-semester upper-level course in human physiology. The course is available to students majoring in Biochemistry, Nutrition, Dietetics, or Neuroscience in either the third or fourth year of their respective undergraduate programs. During this study, the course was organized according to body system and consisted of lectures and laboratory classes. The evaluation of students was comprised of laboratory reports worth 20% of the grade and a series of multiple-choice examinations that were collectively worth 80% of the overall grade. In each semester, there were two midterm examinations consisting of 35 items and a final examination of 80 items. Grades are awarded on a university-wide scale (A: 80–100%, B: 65–79%, C: 55–64%, D: 50–54%, and F: <50%).

The course was delivered by nine faculty members, which included one of the study authors (J. D. Kibble). For the most part, one faculty member delivered lessons about one body system. The Course Director provided faculty members with a syllabus and created a simple examination blueprint that specified the number of items required in each major topic area. The faculty members generated questions individually and were asked to submit them to a coordinator for entry into a central question bank, which was developed using LXRTest 6.1 software (www.LXRTest.com). Questions were organized into organ system folders; additional meta-data for each item included author, year, and topic and subtopic descriptions. Faculty members were asked to assign an intended difficulty score (I = easy, 2 = moderate, 3 =
and 3 = hard) for each item. Faculty members were also asked to use a truncated Bloom’s taxonomy (2) to categorize their items as follows: 1 = knowledge (Does the question only require factual recall?), 2 = comprehension (Does the question require an understanding of a physiological mechanism?), and 3 = application (Does the question require prediction, calculation, data interpretation, or graphical interpretation?). As a quality control measure, J. D. Kibble and the Course Director independently assigned a Bloom’s taxonomy score to all test items and then met to reach a consensus score.

Data analysis. Data reported from each examination included the median, mean, and SD of student scores (in %). Cronbach’s α was calculated as a measure of the internal consistency. Item analysis included the proportion of students answering correctly (hereinafter referred to as the “P value”) and an item discrimination index, which was calculated as a relative point biserial correlation coefficient and tests how closely related is success on a question to success on the test. These data were standard variables obtained from the LXR testing software. A postexamination review was undertaken before final exam scores were confirmed. The review considered items with a negative discrimination index as well as items queried by students via written challenges. The question author and Course Director reached consensus about potentially problematic items, and any found to have flaws were removed from scoring.

The relationship between faculty predictions of item difficulty and actual item difficulty was examined in two ways: 1) Spearman’s ρ correlations using item P values as an interval measure of actual item difficulty and 2) cross-tabs using classifications of item P values as an ordinal measure of actual item difficulty, where items with P values of 0.80–1.00 (i.e., corresponding to a grade of A) were classified as easy, P values of 0.55–0.79 (i.e., corresponding to grades of B or C) were classified as moderate, and P values of 0–0.54 (i.e., corresponding to grades of D or F) were classified as hard. The classification of item P values into ordinal categories allowed the calculation of rates of successful prediction and application of χ²-analysis. Spearman’s ρ correlations were also computed to assess the relationship between faculty classifications of the item cognitive level and J) predicted item difficulty, 2) actual item difficulty, and 3) item discrimination. Cohen’s κ was calculated as a measure of agreement between faculty classifications of the item cognitive level and consensus taxonomic scores. Statistical analyses were completed using IBM SPSS/PASW Statistics 18.0 (IBM, Chicago, IL). A significance level of 5% was used for all statistical tests. This study was approved by the Memorial University Human Investigations Committee, and participants gave informed consent.

RESULTS

Nine faculty members taught the human physiology course, and eight of them gave informed consent to participate in this study. Of the 300 examination questions administered, 272 questions were used for analysis. The 28 items removed from analysis included those from the faculty member not giving consent, 13 items that were not single best answer-type questions, and 4 items that were found to be flawed during the postexamination review. Table 1 shows the summary data for each examination. Test reliability ranged from 0.70 in the midterm examinations to 0.92 for the final examinations at the end of semester. While the difference in Cronbach’s α between tests was significant, it is most likely due to the fact that the final exams had more than double the number of test items than the other exams. The overall mean examination score for the whole course was 75%, which is toward the upper end of the grade B range.

Faculty predictions of item difficulty level. Of 272 exam items, faculty members intended 117 (43%) to be easy, 116 (43%) to be of moderate difficulty, and 39 (14%) to be hard. Compared with the grade-based performance categories, faculty members as a group successfully predicted 130 of 272 items (48%); 57 of 117 easy items (49%), 63 of 116 moderate items (54%), and 10 of 39 hard items (26%). Overall, faculty members’ predictions of item difficulty were modestly but significantly correlated with item P values (r = −0.19, P < 0.01). Figure 1 shows this moderate correlation between predictions of difficulty and actual student scores and also highlights the relative frequency of successful and unsuccessful predication in each difficulty category.

It is interesting to note that the distribution of items across actual difficulty levels was similar to that intended by faculty members, despite the fact that many individual item predictions were incorrect. These data are shown in Table 2 and account for the fact that the overall difficulty of an examination turned out to be appropriate.

Individually, faculty members differed significantly in overall rates of successful prediction (χ² = 16.84, P = 0.02). Overall rates of successful prediction ranged from 27% to 75%; cell sizes were insufficient to allow for additional χ²-analyses by difficulty level. Correlations between individual faculty predictions of item difficulty and actual item difficulty ranged from 0.07 (weaker) to −0.56 (stronger). Figure 2 shows rates of successful prediction across difficulty level by faculty member, and Table 3 shows complete individual and collective prediction data.

Faculty classifications of item cognitive level. Faculty members’ classifications of item cognitive level were most closely related to their predictions of item difficulty (ρ = 0.63, P < 0.01). In general, faculty members classified the majority of easy items at the knowledge level (90 of 117, 77%), moderate difficulty items at the comprehension level (65 of 116, 56%), and hard items at the application level (27 of 39, 69%). Figure 3 shows complete frequency data for the prediction of item difficulty levels by classification of cognitive levels.

To test whether the classification of items by cognitive level could be done consistently, the consensus ratings of J. D. Kibble and the Course Director were also analyzed. The classifications by item writers and the consensus raters matched on only 155 of 272 items (57%); 88 items matched at the knowledge level, 32 items matched at the comprehension level, and 35 items matched at the application level (κ = 0.33, P < 0.01).

In terms of whether the cognitive level of an item was a determinant of the outcome on a test, no significant relationships were observed between faculty members’ classifications of cognitive level and 1) actual item difficulty (ρ = −0.09, P = 0.14) or 2) item discrimination (ρ = 0.05, P = 0.42). Similarly, there were no significant relationships between consensus rat-
ers’ classifications of cognitive level and 1) actual item difficulty ($\rho = -0.04, P = 0.50$) or 2) item discrimination ($\rho = 0.06, P = 0.34$).

**DISCUSSION**

The impetus for this study came from an initiative to start a new question database for multiple-choice items that would facilitate the administration of examinations in an undergraduate human physiology course. In the absence of historical data, we hypothesized that faculty members can accurately predict the difficulty of their questions. If this is the case, it will be a useful annotation for new items entered into the question bank as we strive to set exams that conform to an arbitrary grading scale. We also hypothesized that assignment of a learning taxonomy to our questions would be a predictor of item difficulty. The main findings of the study based on the data were that 1) the ability of faculty members to predict item difficulty is detectable but weak, 2) the assignment of cognitive levels to questions cannot be done with a high degree of consistency, and 3) the item writer’s perception of learning taxonomy has no relationship to actual item difficulty. Although these item annotations proved to be less useful than we hypothesized, there were unanticipated benefits, as discussed below, in terms of revealing aspects of the hidden curriculum.

Our study has limitations in terms of the ability to generalize findings, since it is based on data from eight faculty members teaching in a specific course. Our context will be familiar to many, though: we are a group of relatively autonomous faculty members who are mostly content experts that each delivers a particular section of the course. We work in a research-intensive environment, and there are few opportunities to meet as a whole group to discuss the curriculum or to review examination questions. It is common for the Course Director to send individual requests to faculty members for questions and then collate the examinations herself. This context accounts for our study design in which individual faculty members classified their own items.

Our hypothesis that faculty members could estimate the difficulty of individual items turned out to be statistically correct, but was not very impressive, and is probably not worth pursuing as a means to inform standard setting. Figure 2 shows a wide variation in the ability of individual faculty members to appreciate the difficulty of their questions. Presumably the ability to succeed in this endeavor relates to knowing what
material was taught, understanding the capabilities of students, and having a sense of what they are really learning during the course. Anecdotally, we could not detect any obvious relationship between the ability to predict item difficulty and the length of teaching experience, age, or sex of the faculty member. It is worth pointing out, though, that despite a modest ability to predict individual item outcomes, our examinations had consistent levels of difficulty, which conformed with institutional grading norms and were acceptable to students, faculty, and the administration. Examination reliability was also good (0.70–0.92), particularly for the larger end-of-semester assessments. Therefore, it appears that collective wisdom was at work despite the fact that we were not able to meet as a group of experts ahead of examinations to apply formal standard-setting methods.

Our second hypothesis was that the assignment of a learning taxonomy to questions would provide an objective measure related to difficulty. This expectation was false, since our data showed no relationship between the item cognitive level and student score. This differs from a study (8) of pharmacy students in which explanation-level questions were found to be the most difficult. However, in the same study, there was no difference between knowledge- or prediction-level questions, which does not support the notion of a systematic relationship between cognitive level and item difficulty. This conclusion was supported in a study (5) that classified items on the Medical Council of Canada licensing examination and did not find a relationship between cognitive level and item difficulty (5). In a similar vein, two other studies (11, 13) failed to observe any predictive validity of higher-order questions in

---

**Table 3. Success rates for item difficulty prediction for individual faculty members**

<table>
<thead>
<tr>
<th>Faculty member</th>
<th>Rate of Successful Prediction†</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ρ</td>
<td>%</td>
<td>Number</td>
<td>%</td>
<td>Number</td>
</tr>
<tr>
<td>Faculty member 1</td>
<td>−0.35</td>
<td>64</td>
<td>7/11</td>
<td>77</td>
<td>13/17</td>
</tr>
<tr>
<td>Faculty member 2</td>
<td>−0.14</td>
<td>48</td>
<td>21/44</td>
<td>63</td>
<td>19/30</td>
</tr>
<tr>
<td>Faculty member 3</td>
<td>−0.56</td>
<td>83</td>
<td>5/6</td>
<td>80</td>
<td>4/5</td>
</tr>
<tr>
<td>Faculty member 4</td>
<td>−0.44†</td>
<td>50</td>
<td>6/12</td>
<td>50</td>
<td>8/16</td>
</tr>
<tr>
<td>Faculty member 5</td>
<td>−0.34*</td>
<td>56</td>
<td>10/18</td>
<td>45</td>
<td>9/20</td>
</tr>
<tr>
<td>Faculty member 6</td>
<td>−0.23</td>
<td>50</td>
<td>2/4</td>
<td>67</td>
<td>2/3</td>
</tr>
<tr>
<td>Faculty member 7</td>
<td>−0.07</td>
<td>15</td>
<td>2/13</td>
<td>42</td>
<td>5/12</td>
</tr>
<tr>
<td>Faculty member 8</td>
<td>−0.001</td>
<td>44</td>
<td>4/9</td>
<td>23</td>
<td>3/13</td>
</tr>
<tr>
<td>All faculty members</td>
<td>−0.19†</td>
<td>49</td>
<td>57/117</td>
<td>54</td>
<td>63/116</td>
</tr>
</tbody>
</table>

Rates of successful prediction are presented as rate = (number of successfully predicted items/number of total predicted items). *Significant at the 0.05 level; †significant at the 0.01 level.
terms of their ability to predict later clinical performance in medical students. We should also call into question whether faculty members can reliably assign a cognitive level to multiple-choice items since we found agreement in only 155 of 272 items when comparing the consensus rating with the faculty rating. Others have noted similar difficulty in making reliable judgments about cognitive levels, which may depend on the seniority of the rater. As a result, it is our view that using item classification by learning taxonomy is unlikely to add much to an analysis of student performance. This result surprised us inasmuch as current best practice recommends the creation of examination blueprints, which often include a table of specification that identifies the proportion of higher- and lower-order items (3). Even if this can be done reliably, the outcome data appear to have little meaning.

Perhaps a better reason to invest effort in writing test blueprints and in annotating the questions is to assure that the tests are valid (i.e., test what they purport to). For example, in our course, we emphasize the need for student understanding of physiology, and this is aligned with the evidence that we used 145 of 272 items of comprehension level or above. In making the effort to focus on our tests through this project, we were also rewarded with some useful and unexpected discussion when the data were shared with the faculty members at our official end of course meeting. The discussion around the data revealed aspects of the so-called hidden curriculum, which the data allowed us to better reflect upon. Ideally, there should be good alignment between learning objectives, learning experiences, and evaluation. When this is not the case, students take a variety of cues about what knowledge, skills, and attitudes should be assimilated, and an informal or hidden curriculum develops (7). In a study that surveyed how students used learning objectives, classroom experiences, or a test blueprint to inform how to study, it was found a significant proportion of students relied on cues from outside the formally administered curriculum. Role modeling by faculty members plays a big role in determining the “informal curriculum.” For example, if a professor is passionate about laboratory classes and dismissive of lecture, students may take the cue that the test will emphasize laboratory-based teaching. The hidden curriculum is shaped by institutional policies and practices. For example, if a laboratory class is always scheduled late in the day, or is poorly resourced, this may send a message to students that laboratory-based learning is not valued.

At our faculty meeting, the ability to view the distribution of intended difficulty and cognitive levels of questions resulted in a discussion that revealed a spectrum of philosophies about the purpose of assessment and grading. Some faculty members made a strong connection between learning objectives and evaluation, whereas others argued that the learning experiences were a more important focus. Considering that no major differences were observed between individual faculty members in the distribution of intended difficulty or cognitive levels, it was surprising that faculty members did not have a shared philosophy on assessment. A review of the item analysis data was also a useful catalyst for discussion about our collective philosophy on grading. Some of the faculty members placed high value on the item discrimination index as an indicator of item quality and were more attracted to the idea of norm-referenced grading in which the best performing students in a group are rewarded. Others viewed items more subjectively and seemed better aligned with a criterion-referenced grading philosophy in which all students reaching a benchmark standard are rewarded. We take from this exercise the importance of having such discussions up front to help ensure that a hidden curriculum does not become a driving force for what is learned by students.

In conclusion, teachers have a modest ability to estimate the difficulty of their multiple-choice questions, but, in our view, this ability is not good enough to inform standard setting of examinations. The assignment of learning taxonomies to multiple-choice questions has no relation to the difficulty of ques-

---

**Fig. 3. Relationships between an author’s intended item difficulty and their Bloom’s taxonomy score.**
tions and cannot be used to control examination difficulty. A shared understanding about testing and grading philosophy should not be assumed, and intentional meetings among faculty members on this issue may help to avoid the development of a hidden curriculum.

ACKNOWLEDGMENTS

The authors thank Penny Hansen, the Course Director, for spending the time to classify questions and discuss aspects of this study as well as Moshe Feldman for the initial statistical consultation.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

REFERENCES