Pharmacy Education (2020) 20(1) 303 - 312
https://doi.org/10.46542/pe.2020.201.303312

RESEARCH ARTICLE

Implementation and students' perception of a criterion-referenced standard setting in a therapeutics course

YeeAnn Chen[1], Crystal Zhou[2], Andrew Leeds[2], Jaekyu Shin[2]

[1]University of Michigan, USA. *(University of California, San Francisco at time of research)*
[2]Department of Clinical Pharmacy, School of Pharmacy, University of California, San Francisco, USA.

Keywords
Pharmacy Education
Assessment
Standard Setting
Criterion-references Methods
Student Stress

Correspondence
Jaekyu Shin
Department of Clinical Pharmacy
School of Pharmacy
University of California
San Francisco
jaekyu.shin@ucsf.edu

Abstract

Objective: To implement a criterion-referenced method to set standards for grading written tests in a didactic course and to assess students' perceptions of the implementation. **Methods:** The Angoff method, a criterion-referenced method, was implemented in a therapeutics course with a letter grading system. Students were surveyed on their perceptions of the method including stress and test anxiety level after the course. **Results:** Of 122 students enrolled, 118 responded. More than 60% of respondents felt that the criterion-referenced method was fairer and reflected competency better than a norm-referenced method. The percent of respondents who felt that the new method increased the level of stress and test anxiety was higher than that of those who did not. **Conclusions**: A criterion-referenced method was successfully implemented in a pharmacy didactic course with a letter grading system. The implementation was overall favourably received by students although it may have increased the level of stress and test anxiety.

Introduction

Pharmacy education has been moving toward competency-based education (Katajavuori *et al*., 2017; Koster, Schalekamp, & Meijerman, 2017; Gregory *et al*., 2019; Katoue & Schwinghammer, 2020). In competency-based education, it is important to document that students demonstrate competency on a test at an appropriate level of their progress in the curriculum. In evaluating competency, setting a standard or a passing score on a test in a defendable and reliable manner is crucial as the standard is used to make an important decision - whether the student's performance meets minimum competency (Yudkowski, Downing, & Tekian, 2009).

There are two types of methods used to set standards: norm-referenced and criterion-referenced methods (Bendavid, 2000; Yudkowski, Tekian, & Downing, 2006; Yudkowski, Downing, & Tekian, 2009). Norm-referenced methods determine passing scores through comparing students' scores to a well-defined normative group's scores (e.g., scores of all of the students who took the test). As a result, the student's score on a test using a norm-referenced method is interpreted relative to the normative group's score. On the other hand, criterion-referenced methods determine passing scores using predetermined criteria. The student's score on a test using a criterion-referenced method provides information on how well the student knows or understands the content assessed on the test. Since one of the purposes of tests in competency-based education is to determine whether the student's performance meets the minimum competency predefined by faculty, criterion-referenced methods have

been recommended to be used to set passing scores in competency-based education (Holmboe *et al*., 2010; Royal & Guskey, 2015).

Despite the importance of setting standards by using a method that is defendable and reliable, criterion-referenced methods do not appear to be widely used in didactic courses in pharmacy education. The use of a criterion-referenced method has been reported mainly in examinations on skills such as presentation skills (Alston & Love, 2010; Peeters, Sahloff, & Stone, 2010). Only one study has reported on the use of a criterion-referenced method for written tests in a cardiovascular pharmacotherapy course (Schullo-Feulner, Kolar, & Janke 2015). However, this study did not describe what criterion-referenced method was used and how it was implemented. Instead, it focused on the evaluation of written tests in the course. Therefore, the implementation of a criterion-referenced method for written tests to assess knowledge has rarely been documented in a didactic course in pharmacy education.

Many factors may have contributed to the under-utilisation of criterion-referenced methods in pharmacy education. First, there is a cultural factor. Since norm-referenced methods have been used in education for a long time, educators may not be familiar or comfortable with criterion-referenced methods. Even in medical education, where competency-based education started earlier than in pharmacy education, a criterion-referenced method does not appear to have been fully adopted (Holmboe *et al*., 2010). Second, the grading system may play a role. In pharmacy education in the United States, the type of grading systems used in didactic courses has not been formally reported; however, several sources suggest that a letter grading system (i.e., the A, B, C, D, and F grading system), instead of a pass/fail grading system, is more widely used. For example, in one survey administered to pharmacy schools and colleges in the United States, 66% of the schools responded that they utilised the letter grading system for Advanced Pharmacy Practice Experience courses (Pincus *et al*., 2019). In addition, more than half of directors of pharmacy residency programmes viewed pass/fail grades unfavourably (Pick *et al*., 2013). As a result, most pharmacy schools in the United States have kept a letter grading system. Furthermore, educators may feel that a norm-referenced method fits a letter grading system better as students in the same class are compared against each other in the letter grading system (Bendavid, 2000; Yudkowski, Tekian, & Downing, 2006; Yudkowski, Downing, & Tekian, 2009). Third, there is a time and resource factor. Compared with norm-referenced methods, criterion-referenced methods may take more time and resources to set standards; in general, it requires multiple steps: 1) recruiting and training judges; 2) reviewing and scoring test items by individual judges; and 3) discussing score discrepancies between the judges (Bendavid, 2000; Yudkowski, Tekian, & Downing, 2006; Yudkowski, Downing, & Tekian, 2009). In addition, the time and resources it takes may vary depending on the standard setting method used (Bendavid, 2000; Yudkowski, Tekian, & Downing, 2006; Yudkowski, Downing, & Tekian, 2009).

Since norm-referenced methods have been predominantly used in education, students may not be familiar with criterion-referenced methods. As a result, changing to a criterion-referenced method to set standards may impact students' level of stress and test anxiety. In particular, pharmacy students have a high level of stress and a low level of health-related quality of life (Marshall *et al*., 2008; Hirsch *et al*., 2009; Beall *et al*., 2015). Since stress negatively influences students' wellbeing and tests are a trigger of stress, it is important to understand whether changing from a norm-referenced method to a criterion-referenced method impacts students' stress levels (Marshall *et al*., 2008; Hirsch *et al*., 2009). This information may be useful to address students' stress when a criterion-referenced method is implemented to set standards. Additionally, since the assessment drives students' learning, changing to a criterion-referenced method may influence students' study strategies (Beall *et al*., 2015). It may be helpful for educators to know how students may change their study strategies when a criterion-referenced method is adopted because educators may utilise this information to help students better prepare for a test using a criterion-referenced method.

University of California San Francisco (UCSF) launched a new Doctor of Pharmacy curriculum in autumn, 2018. Compared with the previous curriculum, the new curriculum is competency-based. To prepare for the launch of this competency-based curriculum, a pilot project, which could demonstrate the feasibility of using a criterion-referenced method to set standards for written tests in a didactic course, would be useful. Therefore, the objectives of this study are:

1) to demonstrate successful implementation of a criterion-referenced method to set standards for written tests in a didactic course with a letter grading system;

2) to evaluate pharmacy students on their perception of using a criterion- referenced method in determining a passing score on written tests.

Methods

This cross-sectional study was declared to be exempt from full review by UCSF Institutional Review Board.

The Therapeutics II course

At the time of this study, the Doctor of Pharmacy programme was a professional degree programme with the first three years focusing mainly on didactic education and the final year on experiential training. Therapeutics II was a course offered to the second year Doctor of Pharmacy students at the UCSF. This required six-unit course ran for ten weeks, covering the treatment and management of common cardiovascular diseases such as hypertension, coronary artery disease, heart failure and arrhythmias. In 2018 when this study was conducted, 122 students were enrolled in the course.

Written tests in Therapeutic II

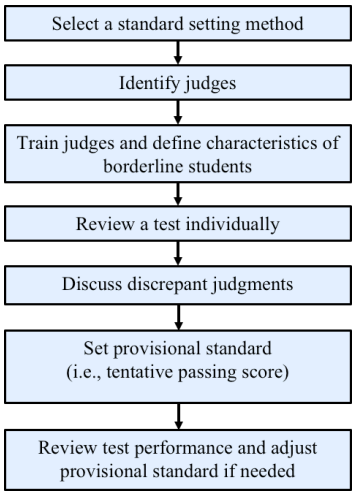
There were four written tests in the course. The authors administered a two-hour long written test every two-three weeks using ExamSoft, a computer-based software (Dallas, TX, USA). Each test included three-four cases and consisted of 26 questions that were mostly multiple-choice type.

The maximum score a student could achieve on each test was 100%. The student's percent test score was calculated with the following formula: the total points that the student actually earned × 100 ÷ the maximum points that could be earned.

Standard setting process

Figure A depicts the standard setting process. The Angoff method with some modifications was chosen because it was widely used and easy to implement (Yudkowski, Downing, & Tekian, 2009). Three judges were selected: the course director, a senior faculty and a junior faculty. All of the judges were instructors of the course with an expertise in cardiovascular therapeutics. In addition, judges had diverse years of teaching experience to minimise over- or under-estimation of student performance due to years of teaching experience. The course director, senior faculty and junior faculty had ten years, more than forty years, and one year of teaching experience, respectively. The rationale behind the course director participating as a judge was to assist the other judges in estimating student performance on each exam question given his seven years of experience as the course director.

Figure A: The standard setting process



The judges met five times. The first meeting was to train judges and to define the borderline student. The other meetings were to set a tentative passing score before each test. Each meeting took about an hour. In the first meeting, the course director trained the other two judges on the overall process of setting standard using the Angoff method. In addition, the judges had a discussion to define 'borderline student' considering the level of competency expected of the second-year pharmacy students at the institution (Searle, 2000). The following was the judges' definition of the borderline student:

> '*A borderline student has variable understanding of a case and concepts and applies information inconsistently. The borderline student knows facts but does not fully understand rationales behind the facts and tends to regurgitate rather than to apply information.*'

After the first meeting, each judge individually reviewed tests and answered the following two questions by using a test review sheet in an Excel file:

- What percentage of borderline students will answer the question correctly?
- Is the question item fundamental or critical to the course content?

In answering the first question, it was emphasised to the judges that they should estimate how well students would perform on each question item instead of how well students should perform. In determining whether the question item is fundamental or critical to the course content, the judges were asked to consider current pharmacy practice.

The first question was used to set a tentative passing point and score, and the second question was used to additionally assess students who had a score in the passing boundary (Searle, 2000). The mean estimated percentage of borderline students who would correctly answer a question item was called a passing point of the question item. To calculate a passing point, the difference in the estimated percentage among the judges must be no more than 20% (Searle, 2000). In the meeting before each test, the judges discussed question items with an estimated percentage difference more than 20%. The following is the detailed score adjustment process. First, the course director compiled judges' estimated percentages on an Excel sheet (Figure B). To make the judges' meeting more efficient, the course director identified question items with an estimated percentage difference of more than 20%. Only these question items were discussed during the judges' meeting. For these items, each judge presented their reasons for their scoring. If there were data on students' performance on similar question items, the course director presented the data. Afterwards, the judges adjusted their estimated percentage, if necessary. Once the estimated percentage difference became no more than 20%, the mean of the estimated percentage was calculated and then this mean was multiplied with points assigned to the question item. The result of these calculations was a passing point of the question item. After calculating a passing point of each question item, all of the passing points were summed. This sum of the passing points was a tentative passing score of a test. As a result, a tentative passing score of a test was set only based on faculty's estimates on students' performance on the test. Of note, for transparency, the authors announced to students a tentative passing score before each test.

To incorporate students' actual performance on the test, a passing boundary was created by using the following formula: a tentative passing score ± standard error of a class mean score.

After a passing boundary was set, whether the student could be considered as passing or not was determined. If the student's score was above the upper boundary, then the student passed the test; if the student's score was below the lower boundary, then the student did not pass the test. If the student's score was within the boundary, then the student must achieve at least 70% of points assigned to the questions determined as fundamental or critical to the course content by the judges in order to be considered as passing (Searle, 2000). Each test had between seven and nine question items that all three judges considered as fundamental or critical to the course content.

Since 70% has been used as the passing point for performance tests such as the Objective Structural Clinical Examination (OSCE) at the School of Pharmacy, the lower boundary was set as 70%. The difference between 70 and the lower boundary was added to the percent score of each student as a score adjustment. This adjusted score was considered as the score for the test and was used to calculate the total course score.

Incorporating a criterion-referenced method into a letter grading system

The Therapeutics II course was letter-graded and the student's total course score came from various assessments including written tests, oral examination, pre-class quizzes, and SOAP notes. The four written tests accounted for 78% of the total course grade and each test

Figure B: An example of the judges' scoring sheet

	% of borderline students to correctly answer					Critical question item?		
Question #	Judge 1	Judge 2	Judge 3	Average (%)	Passing points	Judge 1	Judge 2	Judge 3
1	60	40	75	58.3	2.3	Yes	No	Yes
2	20	20	10	16.7	0.7	No	No	No
3	20	50	40	36.7	1.5	No	No	Yes
4	50	75	80	68.3	2.7	No	No	Yes
5	60	75	90	75.0	3.0	No	No	Yes
6	70	80	50	66.7	2.7	No	Yes	Yes
7	80	80	95	85.0	3.4	Yes	Yes	Yes
8	80	80	95	85.0	3.4	Yes	Yes	Yes
9	60	80	40	60.0	2.4	No	No	Yes
10	90	90	95	91.7	3.7	Yes	Yes	Yes
				Tentative passing score	25.7			

Question items in light blue were not discussed during the judges' meeting because they had a score difference within 20%. Question items in dark blue indicate complete agreement among the judges on whether they were critical to pharmacy practice. The percent average is the mean percent of the judges' scores for the question item. Each passing point was calculated by multiplying the percent average with 4 because each question item is 4 points worth. The tentative passing score is the sum of all of the passing points. Note that this is an example as the actual test contained 26 question items.

had a different weight: test 1 20%, test 2 23%, test 3 27%, and test 4 30% to the total percentage allotted to the written tests. Each subsequent test had a higher weight given the cumulative nature and the intention to motivate students throughout the course.

Students received a letter grade only after they met the following criteria at the end of the course:

1) An adjusted percent test score that was at least 70% or considered as passing on at least two of the four written tests
2) The weighted mean percent score of the four written tests that was at least 70%

In other words, percent scores of non-written tests were considered only after students met both criteria. Students who did not meet both criteria were required to take a written remediation test after the course. If students passed the remediation test, they received a letter grade based on their adjusted four written test score instead of the remediation test score.

Since students' performance on non-written assessments such as oral exams, SOAP notes, and pre-class online quizzes took up 22% of the total course grade, the final course grade varied depending on how each student scored on these non-written assessments. Adjusted percent scores of four written tests were added to the percent scores of the non-written assessments to calculate the total course score in order to determine the final course grade: A if the final course score ≥ 89.5%; B if the final course score was 79.5-89.4%; C if the final course score was 69.5-79.4%; D if the final course score was 64.5-69.4%; and F if the final course score <64.4%.

All students including those who had to take the remediation test eventually received a C or higher grade.

Post-course survey

In the beginning of the course, the students were provided with a written and verbal explanation of the study including the criterion-referenced method. A tentative passing score was announced to the class before each test and informed the students of both upper and lower passing boundaries after each test.

An invitation to survey was sent to all of the 122 students who took the Therapeutics II course in 2018 via Qualtrics after the completion of the course. The survey asked students to rate their perception of fairness, stress/test anxiety level, satisfaction in test score, and likelihood to compare test scores with classmates, all regarding the new criterion-referenced method versus the standard norm-referenced method, primarily using a Likert scale of 'Strongly Agree, Agree, Neutral, Disagree, and Strongly Disagree' (Appendix A). Students also had the opportunity on the survey to explain how knowledge of this new method had impacted their studying, if any, and their overall preference in grading methods. The survey consisted of ten questions and remained open for two months between June and August in 2018. After the initial invitation to participate in the survey, reminder emails were sent to students every week while the survey was open.

Statistical analysis

The authors used descriptive statistics to determine frequency distributions, percentage distributions, means, standard deviations, and inclusive ranges as appropriate. To evaluate an association between the test score and responses to survey questions, a univariable mixed effect linear regression analysis was performed. In this analysis, the test score was treated as the dependent variable, and response to each question and the interaction between the response to each question and time were treated as independent variables. The interaction term was introduced to evaluate if the relationship between the response to the survey question and the test score was different across the time. STATA 16 (STATACorp LLC, College Station, TX, USA) was used and a p-value <0.05 was considered as statistically significant.

Results

Implementation of a criterion-referenced method to set standard

Upon the judges' individual review of each test, the number of question items with a >20% difference in the percentage of borderline students who would correctly answer decreased after Test 1 (Table I). Specifically, about two thirds of the question items on Test 1 had at least 20% difference in the judges' individual estimation whereas a half or less of the question items on Tests 2-4 had. On the other hand, the number of question items the judges disagreed upon whether or not the question item would be fundamental or critical to the course content increased on Test 4 compared with Tests 1-3. Overall, the numbers of question items that the judges needed to discuss in the meeting before each test were 17 on Test 1, 19 on Test 2, 20 on Test 3, and 20 on Test 4, respectively. Each judge spent approximately nine hours for the entire judging process for the four tests.

Table I: Total number of questions with discrepancies in judges' individual estimation, which required discussion during a meeting before each test

Test 1		Test 2		Test 3		Test 4	
Accurate	Critical	Accurate	Critical	Accurate	Critical	Accurate	Critical
17 (65.4%)	15 (58.0%)	12 (46.1%)	16 (61.5%)	13 (50.0%)	16 (61.5%)	13 (50.0%)	19 (73.1%)

*The total number of questions on each test was 26.
Accurate - refers to 'What percentage of borderline students will answer the question correctly?'
Critical - refers to 'Is the question item fundamental or critical to the course content?'

Students' performance on written tests

Table II shows the results of the four written tests. The following number of students did not pass on each written test: 18 on Test 1, 6 on Test 2, 15 on Test 3, and 5 on Test 4, respectively. At the end of the course, only two students did not meet the two criteria to receive a letter grade and were required to take a remediation test after the course.

Table II: The results of the four written tests

	Test 1	Test 2	Test 3	Test 4
Tentative passing point (%)	70.5	63.5	69.5	66.0
Lower boundary (%)	69.5	63.0	68.5	65.5
Standard error of mean of the class score	0.86	0.72	0.77	0.78
Points added to the raw score	0.5	7.0	1.5	4.5

Survey results

Of 122 students, 118 participated in the post-course survey (96.7% response rate). Table III summarises the survey responses. More than 60% of students responded that the criterion-based method was fairer and reflected competency better than the norm-referenced method. Approximately 35-45% of students felt neutral about the following four questions: whether they were more satisfied with the new grading method; whether they preferred the new grading method over the usual method; whether they would like to see the use of the new grading method to be utilised in other courses; and whether they were less likely to compare their test scores with classmates. The proportions of students who agreed or strongly agreed to the first three questions were at least three times higher than those of students who disagreed or strongly disagreed whereas approximately 10% more students disagreed or strongly disagreed to their being less likely to compare their scores with classmates.

Interestingly, the use of a criterion-referenced method appeared to increase the level of anxiety and stress among students because percent of students agreeing or strongly agreeing to increased level of anxiety and stress was approximately 6-10% larger than that of students who disagreed or strongly disagreed.

The use of a criterion-referenced method did not seem to influence students' study behaviour; 62.7% of students reported that it did not change their study strategies to prepare for the tests in the course. Among students who responded that they changed their study strategies, the two most common changes were to practice applying knowledge to cases provided in the course and to form study groups.

Table III: Percentage of students on perception of criterion-referenced methods versus norm-referenced methods

Question item	Strongly Agree or Agree	Neutral	Strongly Disagree or Disagree
Reflected competency better	61.9%	28.8%	9.3%
Fairer	61.0%	32.2%	6.8%
More satisfied	46.2%	39.3%	14.5%
Less likely to compare test scores with classmates	28.2%	35.0%	36.8%
Preferred as grading method	49.1%	44.1%	6.8%
Would like to see utilized in other courses	44.9%	45.8%	9.3%
Increased level of test anxiety	39.8%	26.3%	33.9%
Increased level of stress	43.6%	22.2%	34.2%

Association of survey responses with the test score change

None of the responses to the survey questions was associated with the test score change over time in the course, although there was a trend toward an inverse relationship between the level of test anxiety and test score change (co-efficient, -1.81; 95% confidence interval, -3.95-0.33; p=0.097).

Discussion

In this study, the authors have demonstrated feasibility of implementation of a criterion-referenced method to set standards for written tests in a didactic course with a letter grading system in pharmacy education. Specifically, they successfully used the modified Angoff method and class performance to set passing boundaries for four written tests and incorporated it into the letter grading system. Overall, students received this new method favourably. The majority felt that the new method reflected their competency better and was fairer than the norm-referenced method. However, students seemed to perceive that the new method increased the level of stress and test anxiety. To their knowledge, this is the first study reporting the feasibility and student perceptions of using a criterion-referenced method to set standards for written tests in a didactic pharmacy course.

Previously, one study reported an evaluation of the composition and effectiveness of a criterion-referenced assessment in a cardiovascular pharmacotherapy course (Schullo-Feulner, Kolar, & Janke, 2015). Although this study appeared to set a percent score of 60% as a passing score, it did not describe the method to set this passing score in detail, particularly resources such as faculty judges the study used and ways to incorporate a passing score into a letter grading system. This practical information is important for course directors who are interested in adopting a criterion-referenced method to set standards particularly given that pharmacy education is moving toward competency-based education.

In contrast, this study provides this practical information. It may be more realistic for most pharmacy didactic courses to have a smaller number of judges than eight-ten judges given the limited number of faculty members with expertise in course content who are available in pharmacy schools. One potential concern about using senior verses junior faculty as judges is a bias, which may be introduced by the difference in teaching experience. In previous studies, expert judges were found to be more stringent in setting standards than non-expert judges (Wayne *et al*., 2008; Shulruf *et al*., 2016). In this study, the junior faculty judge had less than a year of teaching experience whereas the senior faculty judge has been teaching for more than 40 years. Interestingly, it was the senior faculty who adjusted the score most often during the meetings of the judges. In total, the senior faculty made adjustments 41 times, which was almost twice more than the junior faculty (23 times). Often, the senior faculty had the highest expectations among the judges. Although the junior faculty judge is an expert in the cardiovascular therapeutics, the high expectations from the senior faculty judge may be in line with the results of the previous studies comparing expert with non-expert judges in setting standards. Since the junior faculty was a student only a couple of years ago, she may have had more realistic expectations than the senior faculty. In addition, the course director's experience with student performance on similar questions in the past was very helpful to set more realistic expectations during the standard setting process.

In this study, each judge spent approximately nine hours for the entire judging process - on average, the time a judge committed to judging was 2.25 hours per test. This amount of time is shorter than the results of previous studies. One study reported that judges spent four hours on average to set standards for clinical evaluation exercise consisting of four cases and six skills areas while judges in another study spent five hours on average for a written test comprising of 100 multiple choice questions (Talente, Haist, & Wilson, 2003; Senthong *et al*., 2013). Since the authors' type of test and the number of questions are different from the tests in these studies (e.g., written verses clinical examination, 25 verses 100 questions), it may be difficult to directly compare the results of this study with those of these studies. The 2.25 hours per test a judge spent in this study may not be a significant barrier to faculty members to participate in the judging process. Passing points can vary depending on the test content and difficulty. Instead of arbitrarily setting them as 70% of the total score achievable on the test, in this study, passing points were set based on faculty judgment and class performance, which made the passing scores more defensible (Searle, 2000). Because of the more rigorous process and defensibility, the majority of the students may have felt that the criterion-referenced method reflected their competency better and more fairly than the norm-referenced method in the post-course survey. Although criterion-referenced methods may fit better for a pass/fail grading system, they may be incorporated into a letter grading system. In this study, a passing point on each test was set at 70%, an equivalent of a C in the letter grading system. Which letter grade a passing point is equivalent, for example a C or a B, depends on the course director and/or school policy. Although there is no data on the use of a letter verses pass/fail grading system in pharmacy schools, the results of previous studies suggest that a letter grading system seems to be more commonly used in pharmacy schools the United States (Pick *et al.*, 2013; Pincus *et al*., 2019). Given that a letter grading system seems to be more commonly used in didactic courses and it may take a long process to change the grading system, this study may provide an example of how to incorporate a criterion-referenced method to set

standards for written tests in a didactic course with a letter grading system. By setting standards in a more defensible way, the course director may be able to ensure that students who have passed the course have the minimum competency.

Previously, it was suggested that the grading system may impact the students' study strategies and habits (Fagin *et al*., 2014). In a study in dental education, students who took the National Board Dental Examination Part 1 for a numerical score spent significantly more time on studying than those taking the exam pass/fail (Fagin *et al*., 2014). In contrast, in this post-course survey, the majority of students reported that they did not change study strategies. Among students who reported changes to their study strategies, the two common changes were to practice applying knowledge to cases provided in the course and to form study groups. These data suggest that students did not study just to pass tests. Instead, it is likely that they studied to achieve a high grade because the course still utilised a letter grading system.

The authors' study also identified that the majority of students had increased their level of stress and test anxiety as a result of the use of a criterion-referenced method. Although the study did not find a significant association between the test score and level of stress or test anxiety, there was a trend toward an inverse relationship between them. Previous studies have reported conflicting results on a correlation between test anxiety and academic performance (Hembree, 1988; Shin & Gruenberg, 2019); however, based on student comments in the authors' study, there seems to be a measure to minimise the extent of the level of test anxiety that may be increased when a criterion-referenced method is implemented. Some students may have been intimidated by the announcement that 'expert' judges set passing scores. In addition, since a tentative passing score was announced before each test, it may have increased students' level of stress and anxiety. For example, a test with a tentative passing score of 60% may be perceived as more difficult than one with 70%, since it implies that the content was deemed complicated enough to where a lower score would be sufficient to 'pass'. When implementing a criterion-reference method, avoiding the term 'expert' and highlighting the process and method as well as keeping a tentative passing score confidential may help minimise an increase in students' level of stress and test anxiety.

There are some limitations to this study. First, since the authors used only the modified Angoff method, the results may not be generalisable to other criterion-referenced methods such as the Ebel method. Second, the course director, who wrote the tests participated as a judge. This could have influenced the other judges' adjustments of their estimation on student performance on test questions. However, the course director was familiar with student performance in the past and there were a limited number of faculty members with expertise in the course content. Third, since the spring quarter when the Therapeutics II course was offered had been historically one of the most stressful quarters in the curriculum due to the amount of the course work, this may have influenced the increased level of student stress. In addition, the Therapeutics II course had four written tests whereas the Therapeutics I course only had two. The increased number of written tests in the Therapeutics II course may have affected the level of test anxiety. Finally, validity and reliability of the survey was not measured, although the survey questionnaire was independently reviewed by two investigators (YC and JS) for clarity multiple times.

Conclusion

The implementation of a criterion-referenced method to set standards for written tests in a pharmacy didactic course with a letter grading system was feasible. In addition, it was received favourably by students although it was associated with an increased level of stress and test anxiety. Given the trend toward competency-based pharmacy education, pharmacy educators who direct a didactic course should consider adopting a more defensible and rigorous way to set a passing score while minimising students' level of stress and test anxiety.

Acknowledgements

This study was supported by the University of California San Francisco School of Pharmacy Troy Daniels Award.

References

Alston, G.L., & Love, B.L. (2010). Development of a reliable, valid annual skills mastery assessment examination. A standardized rubric to evaluate student presentations. *American Journal of Pharmaceutical Education*, **74**(80). https://doi.org/10.5688/aj740580

Beall, J.W., DeHart, R.M., Riggs, R.M., & Hensley, J. (2015). Perceived stress, stressors, and coping mechanisms among Doctor of Pharmacy students. *Pharmacy (Basel)*, **25**,344-354. https://doi.org/10.3390/pharmacy3040344

Bendavid, M.R. (2000). AMEE guide No. 18: Standard setting in student assessment. *Medical Teacher*, **22**, 120-130. https://doi.org/10.1080/01421590078526

Fagin, A.P., Howell, T.H., Da Silva, J., & Park, S.E. (2014). The impact on dental students of changes to the National Board Dental Examination grading system. *Journal of Dental Education*, **78**, 813. https://doi.org/10.1002/j.0022-0337.2014.78.6.tb05734.x

Gregory, D.F., Boje, K.M., Carter, R.A., Daugherty, K.K., Hagemeier, N.E., Munger, M.A., Umland, E.M., & Wagner, J.L. (2019). Leading Change in Academic Pharmacy: Report of the 2018-2019 AACP Academic Affairs Committee. *American Journal of Pharmaceutical Education*, **83**, 7661. https://doi.org/10.5688/ajpe7661

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, **58**(1), 47-77. https://doi.org/10.3102/00346543058001047

Hirsch, J.D., Do, A.H., Hollenbach, K.A., Manoguerra, A.S., & Adler, D.S. (2009). Students' health-related quality of life across the preclinical pharmacy curriculum. *American Journal of Pharmaceutical Education*, **73**, 147. https://doi.org/10.5688/aj7308147

Holmboe, E.S., Sherbino, J., Long, D.M., Swing, S.R., & Frank, J.R. (2010). The role of assessment in competency-based medical education. *Medical Teacher*, **32**, 676-682. https://doi.org/10.3109/0142159X.2010.500704

Katajavuori, N., Salminen, O., Vuorensola, K., Huhtala, H., Vuorela, P., & Hirvonen, J. (2017). Competence-Based Pharmacy Education in the University of Helsinki. *Pharmacy*, **5**(29). https://doi.org/10.3390/pharmacy5020029

Katoue, M.G., & Schwinghammer, T.L. (2000). Competency-based education in pharmacy: A review of its development, applications, and challenges. *Journal of Evaluation in Clinical Practice*, **26**, 1114-1123. https://doi.org/10.1111/jep.13362

Koster, A., Schalekamp, T., & Meijerman, I. (2017). Implementation of Competency-Based Pharmacy Education (CBPE). *Pharmacy*, **5**(10). https://doi.org/10.3390/pharmacy5010010

Marshall, L.L., Allison, A., Nykamp, D., & Lanke, S. (2008). Perceived stress and quality of life among doctor of pharmacy students. *American Journal of Pharmaceutical Education*, **72**, 137. https://doi.org/10.5688/aj7206137

Peeters, M.J., Sahloff, E.G., & Stone, G.E. (2010). A standardized rubric to evaluate student presentations. *American Journal of Pharmaceutical Education*, **74**, 171. https://doi.org/10.5688/aj7409171

Pick, A.M., Henriksen, B.S., Hamilton, W.R., & Monaghan, M.S. (2013). Essential information for mentoring students interested in residency training. *Current Pharmacy Teaching and Learning*, **5**, 546-554. https://doi.org/10.1016/j.cptl.2013.07.017

Pincus, K., Hammond, A.D., Reed, B.N., & Feemster, A.A. (2019). Effect of Advanced Pharmacy practice experience grading scheme on residency match rates. *American Journal of Pharmaceutical Education*, **83**, 6735. https://doi.org/10.5688/ajpe6735

Royal, K.D., & Guskey, T.R. (2015). On the appropriateness of norm- and criterion-referenced assessments in medical education. *Ear, Nose & Throat Journal*, **94**, 252-254. https://doi.org/10.1177/014556131509400701

Searle, J. (2000). Defining competency - the role of standard setting. *Medical Education*, **34**, 363-366. https://doi.org/10.1046/j.1365-2923.2000.00690.x

Senthong, V., Chindaprasirt, J., Sawanyawisuth, K., Aekphachaisawat, N., Chaowattanapanit, S., Limpawattana, P., Choonhakarn, C., & Sookprasert, A. (2013). Group versus modified individual standard-setting on multiple-choice questions with the Angoff method for fourth-year medical students in the internal medicine clerkship. *Advanced Medical Education Practice*, **4**, 195-200. https://doi.org/10.2147/AMEP.S46705

Schullo-Feulner, A., Kolar, C., & Janke, K.K. (2015). A Five-Year Evaluation of Examination Structure in a Cardiovascular Pharmacotherapy Course. *American Journal of Pharmaceutical Education*, **79**(98). https://doi.org/10.5688/ajpe79798

Schuwirth, L.W.T., & van der Vleuten, C.P.M. (2014). How to design a useful test: The principles of assessment. In *Understanding medical education,* (Ed. T. Swanwick) West Sussex, UK: John Wiley & Sons., pp. 243-254. https://doi.org/10.1002/9781118472361.ch18

Shin, J., & Gruenberg, K. (2019). Test-enhanced learning in a pharmacy therapeutics course. *Pharmacy Education*, **19**, 100-107

Shulruf, B., Wilkinson, T., Weller, J., Jones, P., & Poole, P. (2016). Insights into the Angoff method: results from a simulation study. *BMC Medical Education*, **16**, 134. https://doi.org/10.1186/s12909-016-0656-7

Talente, G., Haist, S.A., & Wilson, J.F. (2003). A model for setting performance standards for standardized patient examinations *Evaluation & The Health Professions*, **26**, 427-446. https://doi.org/10.1177/0163278703258105

Wayne, D.B., Cohen, E., Makoul, G., & McGaghie, W.C. (2008). The impact of judge selection on standard setting for a patient survey of physician communication Skills. *Academic Medicine*, **83**(10 Suppl), S17-S20. https://doi.org/10.1097/ACM.0b013e318183e7bd

Yudkowski, R., Downing, S.M., & Tekian, A. (2009). Standard setting. In *Assessment in health professions education,* (Ed R. Yudkowski & S.M. Downing). New York City, NY: Taylor & Francis, pp. 119-148

Yudkowski, R., Tekian, A., & Downing, S.M. (2006). Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teaching and Learning in Medicine*, **18**, 50-57. https://doi.org/10.1207/s15328015tlm1801_11

Appendix A: Survey Questionnaire

1. The new grading method using a faculty panel to set standards reflects my competency better than the usual grading methods (e.g., curving the exam score based on the class performance).
Strongly agree
Agree
Neutral
Disagree
Strongly Disagree

2. The new grading method using a faculty panel to set standards is fair in grading exams compared with the usual methods.
Strongly agree
Agree
Neutral
Disagree
Strongly Disagree

3. I am more likely to accept my exam score with the new grading method than with the usual methods.
Strongly agree
Agree
Neutral
Disagree
Strongly Disagree

4. The new grading method has changed my study strategy to prepare for the exams in Therapeutics II.
Yes
No

5. If you respond to question 4 with "Yes", describe what changes you made to prepare for the exam.

6. The new grading method increased the level of test anxiety compared with the usual methods.
Strongly agree
Agree
Neutral
Disagree
Strongly Disagree

7. The new grading method increased the level of stress compared with the usual methods.
Strongly agree
Agree
Neutral
Disagree
Strongly Disagree

8. I prefer grading my exams with the new grading method over with the usual grading method.
Strongly agree
Agree
Neutral
Disagree
Strongly Disagree

9. I would like to see the use of the new grading method in other courses.
Strongly agree
Agree
Neutral
Disagree
Strongly Disagree