ORIGINAL RESEARCH

# Group versus modified individual standard-setting on multiple-choice questions with the Angoff method for fourth-year medical students in the internal medicine clerkship

Vichai Senthong[1,*]
Jarin Chindaprasirt[1,*]
Kittisak Sawanyawisuth[1]
Noppadol Aekphachaisawat[2]
Suteeraporn Chaowattanapanit[1]
Panita Limpawattana[1]
Charoen Choonhakarn[1]
Aumkhae Sookprasert[1]

[1]Department of Medicine, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand; [2]Central Library, Silpakorn University, Bangkok, Thailand

*These authors contributed equally to this work

Correspondence: Kittisak Sawanyawisuth
Department of Medicine, Faculty of Medicine, Khon Kaen University, Khon Kaen, 40002, Thailand
Tel +66 43 363 664
Fax +66 43 348 399
Email kittisak@kku.ac.th

**Background:** The Angoff method is one of the preferred methods for setting a passing level in an exam. Normally, group meetings are required, which may be a problem for busy medical educators. Here, we compared a modified Angoff individual method to the conventional group method.

**Methods:** Six clinical instructors were divided into two groups matched by teaching experience: modified Angoff individual method (three persons) and conventional group method (three persons). The passing scores were set by using the Angoff theory. The groups set the scores individually and then met to determine the passing score. In the modified Angoff individual method, passing scores were judged by each instructor and the final passing score was adjusted by the concordance method and reliability index.

**Results:** There were 94 fourth-year medical students who took the test. The mean (standard deviation) test score was 65.35 (8.38), with a median of 64 (range 46–82). The three individual instructors took 45, 60, and 60 minutes to finish the task, while the group spent 90 minutes in discussion. The final passing score in the modified Angoff individual method was 52.18 (56.75 minus 4.57) or 52 versus 51 from the standard group method. There was not much difference in numbers of failed students by either method (four versus three).

**Conclusion:** The modified Angoff individual method may be a feasible way to set a standard passing score with less time consumed and more independent rather than group work by instructors.

**Keywords:** Angoff, individual, passing score, standard-setting, multiple-choice questions, internal medicine

## Introduction

Medical educators generally perform various roles, such as attending physicians, teachers, and researchers. As a result, they have limited time for medical education tasks. Time limitation has been shown to be associated with low motivation for educational work.[1]

Setting up the standard passing level is crucial for a licensing examination.[2] The Angoff method is one of the most preferred methods by which to achieve this.[3] Six judges are needed to discuss each test item in short-answer-question-based and extended-matching-question-based papers to increase reliability.[4] In the conventional Angoff method, each judge must rate each item individually but can change their decision at the group deliberation at any time.

The Angoff method has been used to set up standard passing levels since 1971. The working team is composed of a group of judges who each evaluate all test items. The main concept is that the borderline student is able to answer each test item correctly. The Angoff has been modified as a simplified method[4] or three-level method,[5] with both methods being acceptable for setting the passing score. The Angoff was the most popular method for multiple-choice questions during the 1990s.[6] It can be used for both medium- and high-stake examinations, such as licensing examinations,[6–8] and is also appropriate for an Objectively Structured Clinical Examination (OSCE),[9,10] or even testing by computer.[11]

Even though the Angoff method has been proven to be as effective as other methods for determining a standard passing level, such as the whole-test Ebel or Hofstee method,[8,12,13] there are some limitations with the Angoff method. The passing score might depend on the characters and knowledge of the judges.[14–17] Another disadvantage of the method is that it is time-consuming due to the availability requirement of judges for group deliberation. As mentioned earlier, medical educators tend to have time limitations.

The group process and discussion method generally establishes more valid and reliable passing scores than an individual method. The discussion might, however, take a long time and is dependent on the availability of the group members. Therefore, if an individual method gives a similar passing score compared with the group method, using the individual method may be preferable. Logically, the passing scores from an individual method need to be adjusted. This study evaluated the passing scores of the group versus the modified Angoff individual method by using this theory on multiple-choice questions. The aim of this study was to improve the time-consuming process while retaining the advantages of the standard.

## Methods

Six clinical instructors at the Department of Medicine, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand, were invited to participate in the study. All instructors accepted the invitation and were divided into two groups (modified Angoff individual method and group method), matched by the teaching experience.

All instructors were asked to set the standard score for multiple-choice questions that were used for fourth-year medical students of the internal medicine clerkship. The test contained 100 questions with five choices in each question. The Angoff method[3] was introduced to all instructors. Judges were asked to vote "yes" or "no" for each test item as to whether a borderline student would be able to answer the question correctly.

## The modified Angoff individual method

The three instructors in the individual group worked on the test individually. Judges were asked to score each test question in the test with either a "yes" or "no" rating, in terms of its being able to be answered correctly by a borderline examinee. As a substitute for the group discussion, two adjustments were used, including the concordance technique and further adjustment with the reliability index. It is widely agreed upon that using the concordance rating may increase reliability. There were five concordance scores used for adjustment: concordant "yes" score between instructors 1 and 2; concordant "yes" score between instructors 1 and 3; concordant "yes" score between instructors 2 and 3; concordant "yes" score between any two instructors; and concordant "yes" score among three instructors. Three individual passing scores were recorded. In total, there were eight raw numbers. Adjustment of the average of eight raw numbers with a standard error of the mean (SEM) determined the final passing score for the modified Angoff individual method (Table 1). These adjustments were made to substitute for the group discussion of the conventional Angoff method.

Passing scores were calculated by an average of eight raw numbers and reported in four values (value without adjustment, value with adjustment by concordance method, value adjustment with SEM, and value with adjustment by both concordance and SEM). The time each judge took to score all test items (minutes), time to make adjustment by a coordinator (minutes), and time to finish the task after participation in the study (weeks) were also recorded.

## Group method

Group methods followed the conventional Angoff method.[3] The five steps of the Angoff method were outlined to the judges to be followed: (1) judges' discussion of what constituted a borderline examinee; (2) consensus agreement of the borderline examinee; (3) score rating by each judge individually; (4) score recording; and (5) discussion to determine passing score. All instructors had to finish their individual ratings

**Table 1** Adjustment of final passing scores in a modified individual Angoff method by a reliability index

| Reliability index | SEM |
| --- | --- |
| >0.80 | 0 |
| 0.70–0.79 | 1 |
| 0.60–0.69 | 1.5 |
| <0.60 | 2 |

**Note:** The final passing score equals passing score minus standard error of the mean (SEM) of the test, which depended on the reliability index.

before the group discussion. For the final step, the three instructors worked together, discussed each item, and made the final decision for each item as "yes" or "no," representing the passing score for the group method. Judges also discussed their own ratings in regard to the group passing score.

The time each judge used to score all test items individually in minutes, the time for group discussion (minutes), the time for a coordinator to collect data for the group discussion of the final passing score (minutes), the time required for the group discussion meeting (minutes), and time to finish the task after participation in the study (weeks) were recorded.

## Results

The test was used for students in block 1 of the internal medicine rotation at the Faculty of Medicine, Khon Kaen University. There were 94 medical students who took this test. The mean (standard deviation) test score was 65.35 (8.38), with a median of 64 (range 46–82). The Kuder–Richardson Formula 20 was 0.77 with an SEM 20 of 4.06. Cronbach's alpha was 0.77 with an SEM of 4.00, and the reliability coefficient of the whole test was 0.73, with an SEM of 4.57.

The profiles of the instructors in the modified Angoff individual and group methods were comparable in terms of sex, specialty, teaching experience, and academic position (Table 2).

### Passing scores

The eight scores from individual judges and concordant scores from the modified Angoff individual method were 65, 59, 68, 44, 55, 53, 68, and 42 (Table 3). The mean passing score by three judges without adjustment was 64/100 (Table 4).

**Table 2** Profiles of instructor participants in each group

| Variable | Modified Angoff individual method N = 3 | Group method N = 3 |
|---|---|---|
| Male | 2 | 1 |
| Specialty | | |
|   Dermatology | 1 | 1 |
|   Ambulatory medicine | 1 | 1 |
|   Oncology | 0 | 1 |
|   Cardiology | 1 | 0 |
| Teaching experience (years) | 1, 6, 14 | 1, 10, 14 |
| Academic position | | |
|   Clinical instructor | 1 | 1 |
|   Assistant professor | 1 | 2 |
|   Associate professor | 1 | 0 |
| Time spent (minutes) | 45, 60, 60 (mean 55) | 90 |

**Note:** Teaching experience and time spent are shown as individual values for the modified Angoff individual group.

**Table 3** Score for each item by individual instructors and the group

| | Instructors | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Item | 1 | 2 | 3 | 1 and 2 | 1 and 3 | 2 and 3 | Any two | All | Group |
| 1 | Y | Y | Y | Y | Y | Y | Y | Y | N |
| 2 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 3 | Y | Y | N | Y | N | N | Y | N | N |
| 4 | N | Y | Y | N | N | Y | Y | N | Y |
| 5 | Y | N | Y | N | Y | N | Y | N | N |
| 6 | Y | N | Y | N | Y | N | Y | N | N |
| 7 | N | N | N | N | N | N | N | N | N |
| 8 | Y | Y | Y | Y | Y | Y | Y | Y | N |
| 9 | Y | N | Y | N | Y | N | Y | N | Y |
| 10 | Y | N | Y | N | Y | N | Y | N | Y |
| 11 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 12 | N | N | N | N | N | N | N | N | Y |
| 13 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 14 | Y | N | N | N | N | N | N | N | N |
| 15 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 16 | Y | N | N | N | N | N | N | N | N |
| 17 | N | Y | Y | N | N | Y | Y | N | Y |
| 18 | N | Y | Y | N | N | Y | Y | N | Y |
| 19 | N | N | N | N | N | N | N | N | N |
| 20 | Y | N | Y | N | Y | N | Y | N | Y |
| 21 | N | Y | N | N | N | N | N | N | N |
| 22 | Y | N | Y | N | Y | N | Y | N | N |
| 23 | Y | N | Y | N | Y | N | Y | N | N |
| 24 | Y | N | Y | N | Y | N | Y | N | Y |
| 25 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 26 | N | N | Y | N | N | N | N | N | N |
| 27 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 28 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 29 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 30 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 31 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 32 | N | N | N | N | N | N | N | N | N |
| 33 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 34 | N | Y | Y | N | N | Y | Y | N | N |
| 35 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 36 | Y | Y | Y | Y | Y | Y | Y | Y | N |
| 37 | Y | N | Y | N | Y | N | Y | N | N |
| 38 | N | N | N | N | N | N | N | N | N |
| 39 | N | N | Y | N | N | N | N | N | Y |
| 40 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 41 | Y | Y | Y | Y | Y | Y | Y | Y | N |
| 42 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 43 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 44 | N | Y | N | N | N | N | N | N | N |
| 45 | Y | N | N | N | N | N | N | N | N |
| 46 | Y | N | N | N | N | N | N | N | N |
| 47 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 48 | N | Y | Y | N | N | Y | Y | N | Y |
| 49 | Y | Y | Y | Y | Y | Y | Y | Y | N |
| 50 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 51 | N | N | N | N | N | N | N | N | N |
| 52 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 53 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 54 | Y | Y | Y | Y | Y | Y | Y | Y | Y |

(*Continued*)

## Table 3 (Continued)

| Instructors | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Item | 1 | 2 | 3 | 1 and 2 | 1 and 3 | 2 and 3 | Any two | All | Group |
| 55 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 56 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 57 | N | N | N | N | N | N | N | N | N |
| 58 | Y | N | Y | N | Y | N | Y | N | N |
| 59 | Y | N | Y | N | Y | N | Y | N | N |
| 60 | Y | Y | Y | Y | Y | Y | Y | Y | N |
| 61 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 62 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 63 | N | N | N | N | N | N | N | N | N |
| 64 | N | N | N | N | N | N | N | N | N |
| 65 | Y | N | Y | Y | Y | Y | Y | N | N |
| 66 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 67 | Y | Y | Y | Y | Y | Y | Y | Y | N |
| 68 | Y | N | N | N | N | N | N | N | N |
| 69 | N | N | N | N | N | N | N | N | N |
| 70 | N | N | N | N | N | N | N | N | Y |
| 71 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 72 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 73 | N | Y | Y | N | N | Y | Y | N | Y |
| 74 | N | Y | N | N | N | N | N | N | Y |
| 75 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 76 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 77 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 78 | N | Y | Y | N | N | Y | Y | N | Y |
| 79 | Y | Y | N | Y | N | N | Y | N | Y |
| 80 | N | Y | Y | N | N | Y | Y | N | Y |
| 81 | N | N | N | N | N | N | N | N | N |
| 82 | N | N | N | N | N | N | N | N | N |
| 83 | N | N | N | N | N | N | N | N | N |
| 84 | Y | N | Y | N | Y | N | Y | N | N |
| 85 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 86 | Y | N | N | N | N | N | N | N | N |
| 87 | Y | N | Y | N | Y | N | Y | N | N |
| 88 | N | N | N | N | N | N | N | N | N |
| 89 | Y | N | N | N | N | N | N | N | N |
| 90 | N | N | N | N | N | N | N | N | N |
| 91 | N | Y | Y | N | N | Y | Y | N | N |
| 92 | N | Y | N | N | N | N | N | N | N |
| 93 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 94 | N | N | N | N | N | N | N | N | N |
| 95 | N | N | N | N | N | N | N | N | N |
| 96 | Y | Y | Y | Y | Y | Y | Y | Y | N |
| 97 | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 98 | Y | N | N | N | N | N | N | N | N |
| 99 | N | Y | Y | N | N | Y | Y | N | Y |
| 100 | N | Y | Y | N | N | Y | Y | N | Y |
| Score | 65 | 59 | 68 | 44 | 55 | 53 | 68 | 42 | 51 |

**Note:** The row of the instructors means the number of instructors or the agreement of instructors, if indicated; each column shows results of each instructor or agreement of instructors from item number 1–100.
**Abbreviations:** N, no (borderline students are not able to answer the question correctly); Y, yes (borderline students are able to answer the question correctly).

After adjustment with concordance or SEM, the passing score became lower (Table 4). The final passing score in the modified Angoff individual method was 52.18 (56.75 minus 4.57) or 52. The passing score from the group assessment was somewhat lower, at 51 (Table 3). The total number of students

## Table 4 Passing scores by the modified Angoff individual method

| Scoring | Score (/100) |
|---|---|
| Average of three judges | 64.00 |
| Adjusted with concordance | 56.75 |
| Adjusted with SEM | 59.43 |
| Adjusted with concordance and SEM | 52.18 |

**Abbreviation:** SEM, standard error of the mean.

who failed was 4/94 (4.3%) by the modified Angoff individual method and 3/94 (3.2%) by the group method.

## Time spent

The three instructors in the modified Angoff individual method spent 45, 60, and 60 minutes to finish the task individually, while, in the group method, the instructors spent 60, 60, and 90 minutes to finish individual ratings prior to the group discussion. The modified Angoff individual method instructors finished the task within 2 weeks after assignment (1 day, 4 days, and 2 weeks), while the group method took 1 week to make an appointment for the group meeting, which was held 3 weeks later (Table 5). A coordinator spent 60 minutes making the adjustments for the modified Angoff individual method, while the group took 15 minutes to discuss the passing score. The total time spent in all steps was longer in the group method (4 weeks, 5 hours, and 15 minutes).

## Discussion

The standard passing scores decided with the modified Angoff individual method and the Angoff group method were comparable. After adjustments, the modified Angoff individual

## Table 5 Time spent in the modified Angoff individual and group methods

| Variables | Modified Angoff individual | Group |
|---|---|---|
| Time to finish, weeks | 2 (1 day, 2 days, 2 weeks) | 3 |
| Time to make appointment, weeks | 0 | 1 week |
| Time to finish individually, minutes | 45, 60, 60 | 60, 60, 90 |
| Time to finish by group, minutes | 0 | 90 |
| Time to make adjustment, minutes | 60 | 15 |
| Total time spent | 2 weeks and 3 hours 45 minutes | 4 weeks and 5 hours 15 minutes |

**Notes:** "Time to finish" for the modified Angoff individual method was when the last instructor finished individual rating; for the group method, this was when the group discussion was held. "Time to make adjustment" was the time required for a coordinator to perform an adjustment by concordance and standard error of the mean in the modified Angoff individual method, and time taken for group discussion of the final score for the group method. "Total time spent" was the sum of all time spent in each method.

method determined a passing score of 52/100, while the conventional group method resulted in a passing score of 51/100. For the entire process, the modified Angoff individual method was less time-consuming (~2 weeks versus 4 weeks) for establishing the passing scores (Table 5). Number of failed students by both passing scores was comparable.

Instructors who participated in the modified Angoff individual method were happy to do the task individually because they could manage their own schedules. Another benefit of the individual method was that it was less time-consuming. The disadvantage of the individual method is that the instructors had no chance to discuss with other members and the judgment on each item was totally individually dependent. In addition, there was no chance to discuss the final passing score together and compare with the real situation, as was possible in the group method.[3] Judges' discussion tended to increase the passing score, however.[18] The passing score was nonetheless comparable with the standard group method (52 versus 51) after the adjustment with concordance scores and the reliability index.[19,20] The average score of the individual three instructors was quite high (64/100) and lowered after the adjustment by either concordance scores or the reliability index (Table 4).

For the group method, most of the time spent on the administration process was in making an appointment for three busy instructors, amounting to 1 week to decide the date and 3 weeks afterward to bring all three instructors together. The times spent on individually rating the questions were quite similar to those in the individual method, with somewhat higher average times than the individual method (60, 60, and 90 minutes versus 45, 60, and 60 minutes). Similarly, time spent by a coordinator to summarize the final passing score in the group method was not very different from the time spent to adjust the scores in the individual method (15 versus 60 minutes). As mentioned earlier, the group method is more time-consuming due to the difficulty of coordinating the busy schedules of instructors in order that they can have the group meeting. In this study, both groups were comprised of instructors with quite similar teaching experiences and specialties. The final passing scores were somewhat lower than those previously reported in the literature: passing scores are usually around 60/100.[18]

This is a preliminary study comparing modified Angoff individual versus group Angoff methods. The number of judges in each group was lower than the recommended six judges. The number of suitable judges used was actually based on the study for short-answer-question-based and extended-matching-question-based papers, not for a multiple-choice test like in this study.[4] In addition, the number of adjustments by concordance should be at least 720 (6!) scores instead of eight scores. A concordance score is used to increase interrater agreement, while adjustment with the reliability index increases score consistency. Both adjustments were done in this study to substitute for the group discussion of the conventional Angoff method. Further studies should be undertaken to confirm the comparable outcomes by the individual and group methods, comprising three judges in the modified Angoff individual method versus six judges in the group method, or six judges in the modified Angoff individual method (with computer adjustment) versus six judges in the group method.

## Conclusion

This study introduced a modified Angoff individual method, which was done to set the standard passing level for multiple-choice questions. This modified Angoff individual method may be feasible but needs further studies. The modified Angoff individual method was less time-consuming than the conventional group method and better suited the busy working schedules of the instructors.

## Disclosure

The authors report no conflicts of interest in this work.

## References

1. Zibrowski EM, Weston WW, Goldszmidt MA. 'I don't have time': issues of fragmentation, prioritisation and motivation for education scholarship among medical faculty. *Med Educ*. 2008;42:872–878.
2. Downing SM, Lieska NG, Raible MD. Establishing passing standards for classroom achievement tests in medical education: a comparative study of four methods. *Acad Med*. 2003;78:S85–S87.
3. Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL, editor. *Educational Measurement*. Washington DC: American Council on Education; 1971:508–600.
4. Fowell SL, Fewtrell R, McLaughlin PJ. Estimating the minimum number of judges required for test-centred standard setting on written assessments. Do discussion and iteration have an influence? *Adv Health Sci Educ Theory Pract*. 2008;13:11–24.
5. Yudkowsky R, Downing SM, Popescu M. Setting standards for performance tests: a pilot study of a three-level Angoff method. *Acad Med*. 2008;83:S13–S16.
6. Ahn DS, Ahn S. Reconsidering the cut score of Korean National Medical Licensing Examination. *J Educ Eval Health Prof*. 2007;4:1.
7. Hess B, Subhiyah RG, Giordano C. Convergence between cluster analysis and the Angoff method for setting minimum passing scores on credentialing examinations. *Eval Health Prof*. 2007;30:362–375.
8. Yudkowsky R, Downing SM, Wirth S. Simpler standards for local performance examinations: the Yes/No Angoff and whole-test Ebel. *Teach Learn Med*. 2008;20:212–217.
9. Boursicot KA, Roberts TE, Pell G. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Med Educ*. 2007;41:1024–1031.
10. Kaufman DM, Mann KV, Muijtjens AM, van der Vleuten CP. A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Acad Med*. 2000;75:267–271.

11. Siriwardena AN, Dixon H, Blow C, Irish B, Milne P. Performance and views of examiners in the Applied Knowledge Test for the nMRCGP licensing examination. *Br J Gen Pract*. 2009;59:e38–e43.

12. Wayne DB, Barsuk JH, Cohen E, McGaghie WC. Do baseline data influence standard setting for a clinical skills examination? *Acad Med*. 2007;82:S105–S108.

13. Downing SM, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teach Learn Med*. 2006;18:50–57.

14. Wayne DB, Cohen E, Makoul G, McGaghie WC. The impact of judge selection on standard setting for a patient survey of physician communication skills. *Acad Med*. 2008;83:S17–S20.

15. Verhoeven BH, Verwijnen GM, Muijtjens AM, Scherpbier AJ, van der Vleuten CP. Panel expertise for an Angoff standard setting procedure in progress testing: item writers compared to recently graduated students. *Med Educ*. 2002;36:860–867.

16. Boursicot KA, Roberts TE, Pell G. Standard setting for clinical competence at graduation from medical school: a comparison of passing scores across five medical schools. *Adv Health Sci Educ Theory Pract*. 2006;11:173–183.

17. Verheggen MM, Muijtjens AM, Van Os J, Schuwirth LW. Is an Angoff standard an indication of minimal competence of examinees or of judges? *Adv Health Sci Educ Theory Pract*. 2008;13:203–211.

18. Hurtz GM, Auerbach MA. A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educ Psychol Meas*. 2003;63:584–601.

19. Culligan B. *Item Response Theory, Reliability and Standard Error*. Available from: http://www.wordengine.jp/research/pdf/IRT_reliability_and_standard_error.pdf. Accessed June 7, 2013.

20. Langenbucher J, Labouvie E, Morgenstern J. Measuring diagnostic agreement. *J Consult Clin Psychol*. 1996;64:1285–1289.