A Comparison of the Quality of Standard Setting Between Modified Angoff and Bookmark Standard Setting Method

ANUSASANANUN, Sureeporn
SUJIVA, Siridej
Chulalongkorn University, Thailand

Abstract: The purposes of this research are 1) to compare the quality of standard setting between modified Angoff and Bookmark standard setting methods, which includes the validity , reliability , propriety and usability. 2) to investigate the optimal number of judges and the number of items for modified Angoff and Bookmark standard setting methods. A sample of participants involved 12 judges (mathematicians) and 1,074 students in Mattayomsuksa – 3 at a secondary school in Chonburi Province. A mathematical test for assessing examinees consisting of 100 multiple choices items was developed. A computation of difficulty was derived for IRT modle, which provided evidence to justify the use of both standard setting methods. Rating data is compiled by judges using both methods and then the data is analyzed using a method based on Generalizability theory. The expected findings will indicate that the Bookmark method more accurate than the modified Angoff method.

Introduction

There are several standard setting techniques currently in use. One of the most prevalent methods for setting cut scores on these assessments is the Angoff (1971) method of setting standards, it is the method most often used as modified since its introduction (Sireci & Biskin, 1992). The panelists usually begin by drafting descriptions of achievement levels. It typically involves two stages: orientation and training, in the first round of performance estimation, and a second round of performance estimation. In the orientation and training stage judges(panelists) engage in a discussion of the relevant competencies of the target population of examinees for whom the cut score or standard, is to be set. Although widely used, a fundamental flaw of the modified Angoff is the use of item judgment methods to set achievement levels be discontinued. (Shepard, Glaser, Linn, & Bohrnstedt, 1993)

An alternative method for setting cut scores on assessments comprised of selected and constructed response item types has been proposed, called the Bookmark method (Lewis, Mitzel, & Green, 1996). The Bookmark standard setting procedure is a groundbreaking process developed by CTB / McGraw-Hill. Since its inception in 1996, over 28 states have implemented Bookmark to set cut scores on their large-scale assessments. The Bookmark procedure typically includes training, 3 rounds of activities and discussion, and description writing. In addition, it is designed to simplify the judges ' task by a reordered test booklet containing items presented in order of increasing difficulty.

In current situation, most Thai judges normally use modified Angoff method to establish cut scores. The bookmark method has never been used in Thailand. Therefore, this study will contribute to an understanding of Bookmark method that are propriety and usability of methods for Thai judges. An additional consideration in this study is whether the Bookmark or modified Angoff yield more consistent item performance estimates by judges.

Objectives

1) to compare the quality of standard setting between modified Angoff and Bookmark standard setting methods, which includes the validity , reliability , propriety and usability.

2) to investigate the optimal number of judges and the number of items for modified Angoff and Bookmark standard setting methods.

Method

The modified Angoff and Bookmark standard setting method begin with an overview of introduction to standard setting in general. Judges use both of methods to identify seven cut scores separating eight performance levels: excellent, very good, good, almost good, fair, poor, very poor and failure. Twelve mathematic teachers (judges) engage in training and operational modified Angoff and Bookmark methods. After operation, propriety and usability dealing with both of operational methods by interviewing and questionnaire will be use to assess the judges.

Angoff Method

In the modified Angoff method, twelve mathematic teachers is randomly divided into two groups of six. In small group, judges (mathematic teachers) are asked to conceptualize a specific barely master student they had taught. Keeping this student in mind, the judges are a given test item, each judge is asked the question "What percentage of barely master student will answer this item correctly". Repeat for each cut score being set. (good, very good, excellent, poor and very poor). After conducting their initial ratings independently, judges announce their individual item ratings to the entire panel of judges. Following that, frequency of rating data, cumulative percentage of students at each score point and item statistics are shown to the judges. A discussion of their initial ratings with group will be happen. And then, judges are given the opportunity to change their ratings. This percentage ratings are aggregated across items and average across judges to yield the cut score. Facilitator presents each average of cut score from each small group to the large group. The judges discuss their rationals for cut scores again and are given the opportunity to change their ratings independently again. In addition, rating data are aggregated across items and average across judges to yield each cut score and converting each cut score to an IRT scale score.

Bookmark method

In the same way, twelve mathematic teachers is randomly divided into two groups of six. In small group, Each judge receive an ordered item booklet (OIB). The OIB is constructed using items from the test. The items are ordered in terms of difficulty where the easiest item appears first and the hardest item appears last; this ordering is determined by student performance on the test (Buckendahl, Smith, Impara & Plake, 2002 ; Beretvas, 2004). One of benefits of Bookmark procedure is that it can be used for tests of mixed format that include both dichotomous (selected response) and polytomous (constructed response) items. In addition, it is designed to simplify the judges' task by ordering item difficulty and to reduce the number of judgments that judges must make when selecting the final cut score (Mitzel, Lewis, Patz, & Green, 2001).

In round 1 of individual/independent bookmark placements, judges place the almost good bookmark at the first point in the ordered item booklet where they fell that a student who is able to respond successfully to each item up to that point (with at least a 2/3 likelihood) has demonstrated sufficient skills to merit the title "almost good". The criterion probability – termed the response probability (RP)- that is most commonly used with the bookmark procedure is two thirds (Buckendahl, Smith, Impara, & Plake, 2002; Mitzel et al., 2001; Reckase, 2000; Skaggs & Tessema, 2001, repeat for good, very good, excellent, poor and very poor bookmark placements in ordered item booklet.

In round 2 of small group discussion of round 1 results and rerating. Judges discuss the rationale behind their original bookmark placement with other judges at their small group. Following discussion, each judge makes round 2 bookmark placements. Repeat for each cut score being set. (good, very good, excellent, etc.) Then, A small group bookmark placement is calculated for each small group by converting each members' bookmark placement to an IRT scale score, averaging over all small group.

In round 3 of discussion of small group results, the small groups reconvert to large group. The facilitator presents each small groups' bookmark placements and impact data (percent of students expected to fall in each performance level) to the large group. The judges disscuss the rationale behind each small groups' bookmark placements. After that, each judge makes final round 3 bookmark placement, and each bookmark placement is calculated by converting to an IRT scale score, averaging over all large group. In addition, the judges discuss final cut points in order to write performance level descriptions.

Sample

A sample of participants involved 12 judges with at least five years of experience in teaching the mathematics and 1,074 students in Mattayomsuksa – 3 at 12 secondary schools in Chonburi Province**.**

Instrumentation

The research instruments consisted of 1) a mathematical test for assessing examinees consisting of 100 multiple choices items was developed. 2) a rating scale questionaire about standard setting methods for judges was developed. It consisted of explicitness, practicability, implementation, feedback and documentation. (Pitoniak, 2003 ,cited in Cizek; Bunch; Koons, 2004) 3) Guildlines include both of standard setting methods were constructed.

Data Collection and Data Analysis

The data collection consisted 2 phases: 1) Phase 1 data collection with students, the mathematical test has been administered and scored. The data were analyzed item analysis by Item Response Theory using BILOG program. 2) Phase 2 data collection with judges, rating data was compiled by judges using both methods and then the data was analyzed reliablity and convergent validity by Generalizability theory, using GENOVA and comparing the reliability and convergent validity from both methods by test statistics UX_1 's Woodruff and Feldt (1986).

Result and Conclusion

In phase 1, the data from mathematical test was analysis. The finding indicated that most items were almost difficulty and fair discriminant. In phase 2, the expected findings will indicate that the Bookmark more accurate than the modified Angoff standard setting method. Because a ordered item booklet in Bookmark method has been used to remedy the cognitive deficiency of estimating the probability of minimally competent candidate who will answer an item correctly.

References

Beretvas, S. N. 2004. Comparison of bookmark difficulty locations under different item response models. *Applied Psychologcal Measurement. 28*(1): 25-47.

Buckendahl, C. W., Smith, R.W., Impara, J.C. & Plake, B.S. 2002. A comparison of angoff

and bookmark standard setting methods. *Journal of Educational measurement*. *33*(3): 253-263

Cizek, G. J. ; Bunch, M. B. & Koons, H. 2004. Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*. *23*(4): 31-50.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). *Standard setting: A Bookmark approach*. Symposium presented at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ..

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). *The Bookmark procedure: Psychological perspectives.* In G. J. Cizek (Ed.), Setting performance standards (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum.

Reckase, M. D. (2000). Survey and evaluation of recently developed procedures for setting standards on educational tests. In *Student performance standards on the National Asessment of Educational Progress: Affirmation and improvements*. Washington, DC: National Assessment Governing Board.

Shepard, L. A., Glaser, R., Linn, R. L., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement: A report of the National Academy of Education Panel of the evaluation of the NAEP trial state assessment: An evaluation of the 1992 achievement levels. Stanford,* CA: Stanford University, National Academy of Education.

Sireci, S. G. & Biskin, G. H. (1992). Measurement practices in national licensing examination programs: A survey. *CLEAR Exam Review, 3*(1), 21-25.

Skaggs, G., & Tessema, A. (2001, April). *Item disordinality with the bookmark standard-setting procedure***.** Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Woodruff, D. J. and Feldt, L. S. 1986. Tests for eqaulity of several alpha coefficients when their sample estimates are dependent. *Psychometrika*. *51*(3): 393-413.