Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 23 Number 6, May 2018 ISSN 1531-7714

Rounding in Angoff Ratings

Adam E. Wyse, *The American Registry of Radiologic Technologists*

One common modification to the Angoff standard-setting method is to have panelists round their ratings to the nearest 0.05 or 0.10 instead of 0.01. Several reasons have been offered as to why it may make sense to have panelists round their ratings to the nearest 0.05 or 0.10. In this article, we examine one reason that has been suggested, which is that even if panelists are given the opportunity to provide ratings to the nearest 0.01 they often round their ratings to the nearest 0.05 or 0.10 anyway. Using data from four standard settings, we show that in many cases ratings ended in a 0 or 5 when panelists were given the option of using a scale from 0 to 100 in one-point increments and that only about 9% of all ratings ended in a digit other than a 0 or 5. We also examined the impact of different rounding rules and we found that results were quite similar when using different rounding rules. Additional analyses showed the common phenomenon of panelists giving too high of ratings for hard items and too low of ratings for easy items in comparison to conditional p-values. It is suggested that rounding ratings to the nearest 0.05 or 0.10 represent reasonable alternatives to rounding ratings to the nearest 0.01.

Among the methods for determining cut scores on large-scale assessments, the Angoff (1971) standard-setting method is one of the most popular methods (Brandon, 2004; Hurtz & Auerbach, 2003; Plake & Cizek, 2012). In the Angoff method, panelists are asked to review test items and provide item level probability judgments of how they think minimally competent examinees would perform on the items. These item level probability judgments are then analyzed and combined in some way to determine cut scores (see Hurtz & Jones, 2009; Wyse, 2017). Specific implementations of the Angoff method often differ in the number of rounds, the feedback discussed with panelists, the number of different minimally competent examinees for which ratings are collected, the type of items rated, and the rounding rules that panelists use when providing their ratings. The focus of this paper is on the rounding rules that panelists use when providing their ratings.

The rounding rules used when providing Angoff ratings have been the focus of a few research studies. Reckase (2006a) utilized item response theory (IRT)-based simulation methods to investigate the impact of not rounding ratings versus rounding ratings to one or two decimal places. He found small biases when rounding ratings to two decimal places with the largest biases found when cut scores were very high or low on the IRT θ scale. Wyse and Reckase (2012) also explored the impact of rounding using an IRT-based simulation in the context of the National Assessment of Educational Progress (NAEP). Their work showed low bias in cut scores when ratings were rounded to the nearest 0.05 or nearest 0.01, and the potential for very high amounts of bias if ratings were rounded to the nearest whole number. Their findings were consistent with those of Reckase and Bay (1999), who also found that rounding to the nearest whole number can produce high levels of bias depending on the location of the cut score and the distribution of items on the exam. Impara and Plake (1997) performed two studies to compare rounding judgments to the nearest whole number (i.e., the yes/no variation of the Angoff method) versus rounding judgments to two decimal places and found that the two methods produced essentially equivalent cut scores. Plake and Giraud (1998) looked at the impact of rounding ratings to one decimal place versus two decimal places and found some differences between the two approaches. They suggested that a strategy for mitigating differences may be to have people first provide judgments to one decimal place and then revert to providing judgments to two decimal places in the second round.

There are several reasons that are often given as to why it may make sense for panelists to round ratings to the nearest 0.05 or nearest 0.10. One reason stems from the fact that providing Angoff ratings to the nearest 0.01 may be too cognitively complex for panelists. For example, analyses of NAEP standard-setting data have shown that panelists often regress their Angoff ratings in towards the middle of the probability scale and that ratings do not necessarily exhibit high correlations with item p-values (Schulz, 2006; Shepard, 1995; Shepard, Glaser, Linn, & Bohrnstedt, 1994). Similar findings have been reported in Taube (1997), Humphry, Heldsinger, and Andrich (2014), Wyse (2018), and Wyse and Babcock (2018). Simplifying the rating task by having panelists round their ratings at a higher level may make the Angoff rating task less complex (Tannenbaum & Kannan, 2015). Another reason is that it can be logistically easier to have panelists provide ratings that are rounded to the nearest 0.05 or 0.10 because these data take less time to hand enter into a spreadsheet or panelists can be asked to fill out bubble sheets that can be scanned onsite during the meeting (see Cross, Impara, Frary, & Jaeger, 1994; Nichols, Twing, Mueller, & O'Malley, 2010; Plake & Giraud, 1998). In fact, many testing organizations, if they don't utilize computer applications to collect Angoff data, have panelists round ratings to streamline the data collection process. A third potential reason to suggest rounding to the nearest 0.05 or 0.10 is that previous simulation research (see Reckase 2006a; Wyse & Reckase, 2012) seems to indicate that these types of rounding may not dramatically change cut scores. Plake and Giraud (1998) offered a fourth reason, which is that even if panelists are given the option of rounding their ratings to nearest 0.01 they may not use the whole rating scale and end up rounding their ratings to the nearest 0.05 or 0.10 anyway. However, research to support the fact that panelists may round their ratings to the nearest 0.05 or 0.10 even if they are given the option of rounding their ratings to nearest 0.01 has not been presented in the literature on the Angoff method. Investigating whether panelists restrict their ratings in this way is important because it may provide further evidence of challenges faced in implementing the Angoff method when rounding ratings to the nearest 0.01.

In this article, we use data from four credentialing program standard settings to examine the extent to which panelists rounded their ratings to the nearest 0.05 or 0.10 when they were given the option to round ratings to the nearest 0.01. We also explore how cut scores and results may change if ratings were rounded to the nearest 0.05 or 0.10. Our specific research questions are:

1. To what extent do panelists round their ratings to the nearest 0.05 or 0.10 if they were given the option of rounding their ratings to the nearest 0.01?
2. How would cut scores and results change if ratings were rounded to the nearest 0.05 or 0.10?

Based on Plake and Giraud (1998), we expect to find many ratings ending in a 0 or 5 and few ratings ending in other digits. We also expect to find similar results if ratings were rounded to the nearest 0.05 or 0.10 since previous simulation research seems to suggest that these rounding rules tend to have a small impact.

Data and Methods

Data for this study came from four standard settings performed for medical imaging credentialing programs. The four standard settings took place at separate times over a roughly five-year period. The disciplines for the credentialing programs were distinct, but some of the topics assessed on the exams were similar. In particular, the exams contained content related to patient care, safety, image production, and medical imaging procedures. Table 1 provides a summary of the number of rated items and the number of panelists for each Angoff standard-setting study. The number of rated items ranged from 146 to 200 items, while the number of panelists ranged from 9 to 12

Table 1. Summary of Four Different Standard-Setting Studies

Study	Number of Items Rated	Number of Panelists
1	200	9
2	200	12
3	146	10
4	174	11

panelists. The number of rated items and the number of panelists used in the standard settings are typical of credentialing programs.

Each of the standard-setting meetings was facilitated by a staff psychometrician and a subject matter expert in the discipline of the exam. The standard-setting meetings consisted of a half-day of training and a half-day of providing Angoff ratings. The training included a discussion of the definition of the minimally competent examinee, instructions on how to provide Angoff ratings, review of the exam and content specifications, an explanation of how cut scores were calculated from the ratings, and practice providing ratings for a sample of items. The instructions asked panelists to consider a group of 100 minimally competent examinees and estimate how many of them would be able to answer the question correctly. Panelists were told that their item ratings could range from 0 to 100 in one-point increments. These types of instructions are common with the Angoff method. Different than some implementations of the Angoff method, the process only consisted of a single round with no discussion and feedback. The decision to use only a single round was mainly for scheduling and logistical reasons. We divided all ratings by 100 in subsequent analyses.

We used a simple analytical approach to investigate our research questions. First, we figured out what percentage of ratings ended in a 0, 5, or another digit for each panelist in each of the four standard settings. If Plake and Giraud's (1998) assertion is true, one would expect to find that the percentage of ratings ending in a digit other than a 0 or 5 would be low and close to zero for many, if not all, of the panelists. After examining the percentage of ratings ending in different digits, we then rounded ratings to the nearest 0.05 or 0.10 and we examined how the cut scores on the Rasch ability scale under these rounding schemes compared with those when there was no rounding of ratings. To compute the Rasch cut scores, we summed the ratings for each panelist and we translated this sum to the Rasch ability scale through the Rasch test characteristic curve. We found the group cut scores by taking the average of the individual panelist cut scores.

To evaluate the quality of the Angoff ratings, we created scatter plots of the average item ratings versus conditional p-values and we calculated the correlations between the average item ratings and conditional p-values as well as the ratios of the standard deviation of the conditional p-values over the standard deviation of the average item ratings. The scatter plots provide simple graphical displays of the average item ratings versus conditional p-values with the desire being that the points in the plots are close to and randomly distributed above and below the identity line. Such a plot indicates that the average item ratings are close to the values that would be predicted based on the Rasch model and the estimated cut scores.

The correlations provide a measure of the linear association between the average item ratings and the conditional p-values based on the Rasch model. The correlations were estimated as:

$$r_{P(\theta)p} = cor[P_i(\theta), p_i], \qquad (1)$$

where p_i is the average Angoff rating for item *i* and $P_i(\theta)$ is the conditional p-value for item *i* (Clauser et al., 2013; Goodwin, 1999). The conditional p-value was computed as:

$$P_i(\theta) = \{1 + exp[-(\theta - b_i)]\}^{-1}, \qquad (2)$$

where θ is the average group cut score on the Rasch ability scale and b_i is the estimated Rasch difficulty for item *i*. When Equation 1 is more highly positive it implies that panelists were more consistent in terms of how their ratings corresponded with the conditional p-values.

The ratio of the standard deviation of the conditional p-values over the standard deviation of the average item ratings provides a measure of the extent to which panelists may be regressing item ratings in towards the middle of the probability scale and giving ratings for hard items that are too high and giving ratings for easy items that are too low in comparison to the conditional p-values (see Wyse & Babcock, 2018). The standard deviation ratios were estimated as:

$$SD_p = \frac{\sigma[P_i(\theta)]}{\sigma(p_i)}, \qquad (3)$$

where $\sigma[P_i(\theta)]$ is the standard deviation of the conditional p-values and $\sigma(p_i)$ is the standard deviation of the average item ratings. Ideally, Equation 3 should equal 1 with values greater than 1 indicating more variability in the conditional p-values than the average item ratings and typically that panelists have regressed some of their ratings in towards the middle of the probability scale (see Wyse & Babcock, 2018). In terms of the impact of rounding, one would like to see similar or improved correlations and standard deviation ratios when ratings were rounded, and ideally both measures would be close to 1.

Results

Figure 1 displays the percentage of different types of ratings for the four standard-setting studies. Consistent with the hypothesis of Plake and Giraud (1998), we found that many panelists rounded their ratings to the nearest 0 or 5 and gave very few ratings that ended in other digits. In fact, in only two cases did we observe a panelist with more than 50% of their ratings ending in a digit other than a 0 or 5 (panelists 9 and 10 for study number 4), and in only five cases did ratings ending in a digit other than a 0 or 5 represent the highest percentage of ratings provided (panelist 12 for study 2, panelists 2 and 7 for study 3, and panelists 9 and 10 for study number 4). In fact, across all four standard-setting studies we only observed 11 out of 42 panelists with more than 10% of their ratings ending in a digit other than a 0 or 5, and we observed 25 out of 42 panelists with less than 1% of their ratings ending in a digit other than a 0 or 5. In total across all panelists, around 9% of the ratings ended in a digit other than a 0 or 5. These findings confirm that in these four standard-setting studies a majority of panelists tended to not use the whole rating scale and often gave ratings ending in a 0 or 5.

Given the number of ratings that ended in another digit, we do not expect large changes in cut scores or the quality of ratings if ratings were rounded to the nearest 0.05 or 0.10 instead of 0.01. Table 2 bears out these findings and shows that rounding to the nearest 0.05 and 0.10 had little impact on the average group cut score, correlations, and standard deviation ratios. In fact, across the four studies the group cut score and the group correlations between panelists' average item ratings and conditional p-values never changed by more than 0.01 when comparing no rounding to rounding to the nearest 0.05 or 0.10. The standard deviation ratios exhibited slightly larger changes when ratings were not rounded compared to being rounded, but the changes were still

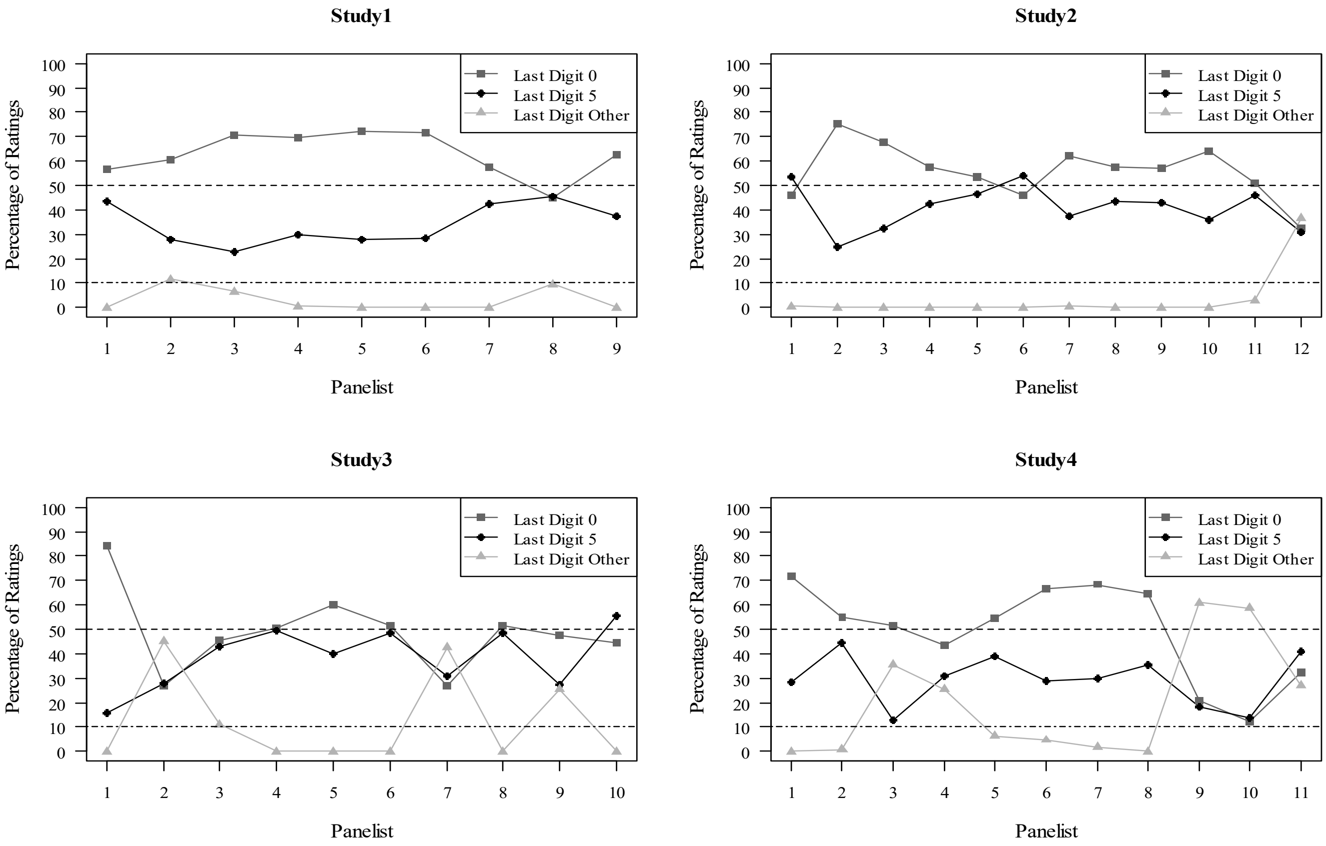

Figure 1. Percentage of Different Types of Ratings for Four Different Standard-Setting Studies
Note. The dotted line shows the threshold for 50% of ratings and the dashed line shows the threshold for 10% of ratings.

Table 2. Average Rasch Cut Scores, Group Correlations, and Group Standard Deviation Ratios When Rounding to Different Digits

	Average θ Cut Score			Correlation			Standard Deviation Ratio		
Study	No Rounding	Rounding to Nearest 0.05	Rounding to Nearest 0.10	No Rounding	Rounding to Nearest 0.05	Rounding to Nearest 0.10	No Rounding	Rounding to Nearest 0.05	Rounding to Nearest 0.10
1	0.64	0.64	0.63	0.56	0.56	0.56	2.51	2.51	2.44
2	1.66	1.66	1.66	0.54	0.54	0.53	2.78	2.78	2.71
3	0.89	0.89	0.89	0.65	0.65	0.63	2.61	2.60	2.53
4	1.04	1.05	1.03	0.77	0.77	0.77	2.05	2.21	2.21

small and less than 0.20. The correlations and standard deviation ratios suggest that the group of panelists generally gave ratings that displayed a similar ordering as the conditional p-values, but the panelists often restricted the range of their ratings and gave ratings that were too high for harder items and too low for easy items. Figure 2 shows the scatter plots of the average item ratings versus conditional p-values when there was no rounding of ratings. These plots confirm the numerical results found with the correlations and standard deviation ratios. In particular, one can see clusters of points with positive slopes and ratings for harder items (i.e., the items towards the left in the plots) that were too high in comparison to the conditional p-values (i.e., above the identity line) and ratings for easier items (i.e., the items towards the right in the plots) that were too low in comparison to the conditional p-values (i.e., below the identify line). Scatter plots when ratings where rounded to the nearest 0.05 and 0.10 were very

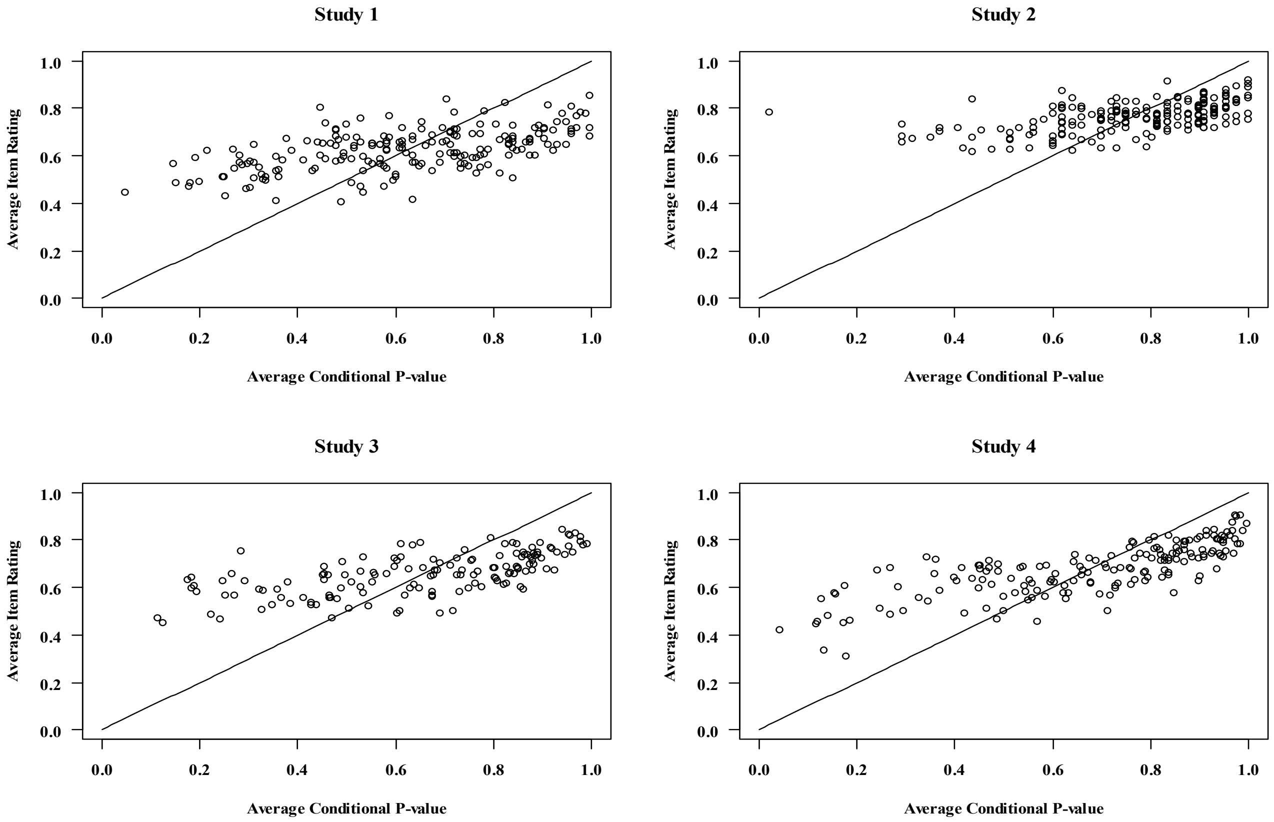

Figure 2. Scatter Plots of Average Item Ratings Versus Average Conditional P-values With No Rounding of Ratings

similar to those found when there was no rounding of ratings.

It could be that the results presented in Table 2 are largely a function of the fact that few panelists gave ratings that ended in a digit other than a 0 or 5, and hence the level of change in the group cut scores and statistics may not be sensitive to rounding to the nearest 0.05 or 0.10. To test this hypothesis, we looked at the individual panelist results to see if there were greater changes at the individual panelist level. The maximum change we observed in a panelist's cut score on the Rasch ability scale when rounding to the nearest 0.05 was 0.02 and the maximum change we observed in a panelist's cut score when rounding to the nearest 0.10 was 0.03. We also observed small changes in individual panelist correlations and standard deviation ratios. These results suggest that even at the individual panelist level the changes tended to be small when rounding ratings to the nearest 0.05 or 0.10.

Discussion and Conclusion

Several reasons have been offered in the literature to support having panelists round their Angoff ratings to the nearest 0.05 or 0.10. Common reasons include that rounding ratings to the nearest 0.05 or 0.10 can save time and help streamline the data collection process, can simplify the cognitive complexity of the Angoff rating task, and does not typically produce large changes in estimated cut scores. Panelists may also round their ratings to the nearest 0.05 or 0.10 even if given the option of rounding their judgments to the nearest 0.01. Our data and results do not directly speak to the first and second reasons for rounding judgments to the nearest 0.05 or 0.10, but they do offer some insight into the other two reasons that have been offered. First, our analyses suggested that many panelists when given the option of rounding their ratings to the nearest 0.01 often gave very few ratings that ended in a digit other than a 0 or 5. Second, our analyses showed that if we rounded ratings to the nearest 0.05 or 0.10 that this had little impact on cut scores, the correlations between item ratings and conditional p-values, and the standard deviation ratios. Our results also add to a growing amount of research, which shows that panelists often restrict the range of their ratings and give ratings that are too high for hard items and too low for easy items in comparison to estimated conditional p-values based on their cut scores and the IRT models used to score the exams.

Of course, an important question to ask when considering the results of this study is how representative our findings of other implementations of the Angoff method. There are some key differences in our implementations of the Angoff method versus other implementations of the method. Most notably, the four studies included in our analyses only consisted of a single round without discussion or feedback. The use of multiple rounds with discussion and feedback is common with the Angoff method. That being said, the first round of ratings and the training and instructions given to panelists were very representative of typical implementations of the Angoff method. That is, it is common to explain the Angoff rating process to panelists, talk about the exam, have panelists practice providing ratings, discuss the definition of the minimally competent examinee, and ask panelists to consider a group of 100 minimally competent examinees and round their ratings to the nearest 0.01 when providing ratings. Hence, we believe our results well-represent rounding in the first round of an Angoff standard setting and seem to indicate that panelists often give a large percentage of ratings that end in a 0 or 5 even if given the opportunity to provide ratings that end in other digits. We also expect that the impact of rounding to the nearest 0.05 or 0.10 will not be large unless there are other oddities in the panelists' ratings, such as a panelist giving many harsh or lenient ratings or further restricting their use of the rating scale such that ratings are predominantly rounded up or down.

There are a few important practical recommendations that can be derived from this work. First, it seems that having panelists round their ratings to the nearest 0.05 or 0.10 is a viable option when performing the Angoff method given that we found that these rounding rules had very little impact on cut scores or the quality of ratings and panelists often rounded their ratings to these levels anyway. In fact, recent standard settings at the credentialing organization that oversees the four credentialing programs investigated in this study have instructed panelists to round their ratings to the nearest 0.05. Second, the results of this study continue to point to the need look at plots and statistics that may capture panelists' tendency to give too high of ratings for hard items and too low of ratings for easy items in comparison to conditional p-values. These rating patterns are often observed with the Angoff method no matter the rounding rule applied. It is important to be aware of and investigate these rating patterns since these rating patterns can influence cut score estimates (see Reckase, 2006b; Wyse, 2017) and the validity of Angoff standard-setting results.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, 17, 59-88.

Clauser, B. E., Mee, J., & Margolis, M. J. (2013). The effect of data format on integration of performance data in an Angoff standard-setting exercise for a medical licensing examination: An experimental study. *International Journal of Testing*, 13, 65-85.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on national teacher examinations. *Journal of Educational Measurement*, 21, 113-129.

Goodwin, L. D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. *Applied Measurement in Education*, 12, 13-28.

Humphry, S., Heldsinger, S., & Andrich, D. (2014). Requiring a consistent unit of scale between the response of students and judges in standard setting. *Applied Measurement in Education*, 27, 1-18.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to Angoff method on cutoff scores and judge consensus. *Educational and Psychological Measurement*, 63, 584-601.

Hurtz, G. M., & Jones, J. P. (2009). Innovations in measuring rater accuracy in standard setting: Assessing "fit" to item characteristic curves. *Applied Measurement in Education*, 22, 120-143.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, 34, 353-366.

Nichols, P. D., Twing, J., Mueller, C. D., & O'Malley, K. (2010). Standard-setting methods as measurement processes. *Educational Measurement: Issues and Practice,* 29(1), 14-24.

Plake, B. S., & Giraud, G. (1998, April). Effect of a modified Angoff strategy for obtaining item performance estimates in a standard setting study. Paper presented at the Annual Meeting of the American Educational Research Association. San Diego, CA.

R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. URL http://www.R-project.org .

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Denmark Paedogogiske Institut.

Reckase, M. D. (2006a). A conceptual framework for a psychometric theory of standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice*, 25(2), 4-18.

Reckase, M. D. (2006b). Rejoinder: Evaluating standard setting methods using error models proposed by Schulz. *Educational Measurement: Issues and Practice*, 25(3), 14-17.

Reckase, M. D., & Bay, L. (1999, April). Comparing two methods for collecting test-based judgments. Paper presented at the annual meeting of the National Council on Measurement in Education. Montreal.

Schulz, E. M. (2006). Commentary: A response to Reckase's conceptual framework and examples for evaluating standard setting methods. *Educational Measurement: Issues and Practice*, 25(3), 4-13.

Shepard, L. A. (1995). Implications for standard setting of the national academy evaluation of the national assessment of educational progress achievement levels. In Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessment (Vol. 2, pp. 143-160). Washington, DC: National Assessment Governing Board and National Center for Educational Statistics.

Shephard, 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 on the evaluation of the NAEP trial state assessment: An evaluation of the 1992 achievement levels. Stanford, CA: National Academy of Education.

Tannenbaum, R. J., & Kannan, P. (2015). Consistency of Angoff-based standard-setting judgments: Are item judgments and passing scores replicable across different panels of experts? *Educational Assessment*, 20, 66-78.

Taube, K. T. (1997). The incorporation of empirical item difficulty data into the Angoff standard-setting procedure. *Evaluation & the Health Professions*, 20, 479-498.

Wyse, A. E. (2017). Five methods for estimating Angoff cut scores with IRT. *Educational Measurement: Issues and Practice*, 26(4), 16-27.

Wyse, A. E. (2018). Regression effects in Angoff ratings: Examples from credentialing exams. *Applied Measurement in Education*, 31, 68-78.

Wyse, A. E., & Babcock, B. (2018). A method for detecting regression of hard and easy item Angoff ratings. Manuscript submitted for publication.

Wyse, A. E., & Reckase, M. D. (2012). Examining rounding rules in Angoff-type standard-setting methods. *Educational and Psychological Measurement*, 72, 224-244.

Citation:

Wyse, Adam E. (2018). Rounding in Angoff Ratings. *Practical Assessment, Research & Evaluation*, 23(6). Available online: http://pareonline.net/getvn.asp?v=23&n=6

Author's Note

The author would like to thank Ben Babcock and Lauren Wood for comments they provided on an earlier version of this manuscript. Any opinions, findings, conclusions, or recommendations expressed in this manuscript are those of the authors and are not necessarily the official position of The American Registry of Radiologic Technologists.

Corresponding Author

Adam E. Wyse
The American Registry of Radiologic Technologists
1255 Northland Dr.
St. Paul, MN 55120

email: adam.wyse [at] arrt.org