



Some Guidance on Proposal Reviewing for the 2023 NCME Annual Meeting

THANK YOU for volunteering to review proposals for the 2023 NCME Annual Meeting! We appreciate your willingness to donate your time and expertise in service of making the program interesting and impactful for all attendees.

Proposal reviewing differs from rating constructed response items or essays, in that there are no criteria for a “correct” answer and it’s hard to have a well-defined rubric when there is no common prompt. We hope the following information will be useful in trying to ensure all raters use a common scale, or perhaps, a more similar scale is more realistic.

We do want to make clear that we are asking multiple reviewers to rate each proposal, and that we, as the Program Committee, also look at each proposal as well as the comments and ratings you all provide. So please do not agonize that you may singlehandedly ruin someone’s career if you don’t give their proposal all 5’s, or that you will ruin the conference by letting a flawed proposal through if you don’t give all proposals all 1’s. Note there are also comment boxes, where you may provide information to us, and to the author(s). If you think a proposal is, for example, on an interesting topic, but the design is weak, a comment such as “this would make a good presentation if the authors would heed the suggestions in the comment box” is very helpful. We can certainly communicate with authors and discuss with them about adding an analysis to a study or a panelist to a discussion. In addition to the ratings, comments, and our own review, we also try to balance the content of the program for the diverse interests of the attendees. For example, if there are 57 CAT proposals, some of them will likely be rejected, even if they all receive maximum ratings, to allow for proposals on other topics to also appear in the program.

There are four aspects we are asking you to rate a proposal on, each with a scale of 1 (Strongly disagree) to 5 (strongly agree).

1. The research offers a novel and well-articulated contribution to measurement theory and/or practice.

Suppose we submit a proposal comparing mode effects on a test for high schoolers or an analysis on subgroup differences on a reading test. You might rate them both a “1” as they don’t seem to be adding much to the existing literature. However, maybe we include some new methodology to compute the differences, or we have a novel way to present findings that seems like it would be appealing, or we dig into process data in new ways. Maybe we are looking at a non-cognitive skill area that has not been widely studied. Or maybe we provide concrete suggestions on how to minimize differences—and perhaps we even have empirical data to test them out. The later aspects would likely receive a higher rating.

If we submit a proposal stating our opinion on something, for example, test optional is good (or bad), that likely should receive a low rating unless there is something like new data to show the effects over time, or some basis for our opinion. A panel of people with different viewpoints who are providing opinions based on their roles (say, college admissions officer, teacher of entry level course,

college counselor trying to avoid student dropouts), might rank higher on the contribution scale by offering different views and putting them in context.

The contribution does not need to be on the magnitude of introducing a new test theory to earn a 5. Nor should something earn a 5 just because you have not seen it before—a new data model with no practical utility might also rate low.

2. The research methods are well articulated and appropriate.

We realize this may be somewhat difficult to judge in some instances, given the short word requirement, but you should be provided with enough information to make a reasonable judgement. Consider again our proposal on mode differences. If we take convenience samples of whomever happened to test in a mode, and computed means on each group for the comparison, that probably deserves a 1. Existing research with randomly assigned groups, or intact groups with covariates likely provide much better data.

However, there certainly may be cases where existing data is all we have—for example, maybe we are comparing people who choose to test in a testing center versus at home with remote proctoring, but we have novel analyses of process data, or some other innovative information. The research methods should support the question the research is intended to address—one good analysis may rate higher than multiple trivial or marginal ones.

Please also consider “research methods” broadly—a demonstration or panel discussion may not have the same type of research methods as a simulation study on CAT, but the same thought process holds—in your opinion, is the format, the panelists, whatever the proposal consists of appropriate for the goal of the session? A panel discussion on a topic of test optional should have panelists familiar with the topic and represent a variety of opinions, likely focused on addressing specific issues such as, do test optional admittees graduate at the same rate as those admitted including test scores; what supports are in place under both systems, how are students referred to supports, and so on. Obviously, there are multiple ways a panel on this, or any topic, can be configured. But the ‘methods’ aspect evaluates if there is an appropriate plan in the proposal that will support the purpose of the session.

If you as a reviewer have suggestions on how to make the proposal stronger, please provide your thoughts in the comments box.

3. There is evidence that the work is well-defined in scope and will be completed by March 2023.

What we are trying to assess here is if the proposal is accepted, whether it is likely to be a presentation where results from the analyses considered in the second item above are available for presentation.

If the proposal happens to include full results, a proposal will likely score high on this item assuming there is cohesion to them (lots of random or unrelated analyses, even if complete, should probably not rate high). When evaluating a proposal without full results, you may look to see whether key elements are available to ensure the desired results can be obtained on time. For example, for a panel discussion whether all panelists have agreed to participate; for research papers, whether data are already in hand or likely to be available in time to conduct the specified analyses; for simulation

studies, whether data simulation programs are ready or in your judgement are likely to be ready in the timeframe needed. A proposal with a data collection plan involving contacting people in March 2023 would likely rate lower on this item, as might a data simulation with so many conditions it seems unlikely there would be time to complete them.

4. The proposal addresses a topic that would be of high interest to NCME members, and the presentation is likely to be well-attended at the meeting.

Here we are asking for you to provide your best judgment on the level of interest NCME attendees would have for the particular proposal. A proposal does not need to be of interest to everyone to score high. Here are a few example scenarios:

- If you think a proposal on setting cut scores on a multi-stage test for certification would be of high interest to those attendees who are involved with multi-stage certification testing, rate it high.
- If you think it would only be of interest to a narrow slice of NCME attendees, feel free to provide a comment to that effect “This would be of extreme interest but likely to a narrow subgroup”.
- If you think the proposal is not likely to draw an audience, rate it lower on this item.

Note that many factors, such as day and time of a session, what competing sessions happen at the same time slot, size of a room, unexpected things like running into a colleague or needing to get some food or an urgent call from the office all affect the actual attendance at a session. In our quest to make the program as informative and interesting as we can, we want to include those proposals that you believe the attendees will value. Maybe the research methods rating is not at the top, but the topic is so provocative it will interest members and spark debate regardless (especially if you offer comments to improve their methodology). Or maybe it’s a proposal ranking high on other items, but it’s unlikely to interest more than the presenters’ close friends.

We hope this information provides a bit of guidance on how we hope you interpret the items you are being asked to rate. If you are reviewing a demonstration proposal, the evaluation criteria are slightly different, but the guidance can be applied similarly. Please feel free to reach out to us with comments in the comment box or with questions at NCMEProgramChairs@talley.com.

THANK YOU for being willing to be a reviewer!

Sincerely,

Dongmei Li, Wei Tao, and Alexis Oakley
Co-Chairs, 2023 NCME Annual Meeting Program Committee