Reviewers' Comments to the Authors:

Reviewer: 1

Comments to the Author

I am happy to see papers on G-theory, because, in my view, researchers and practitioners have not fully taken advantage of what it can offer. I appreciated the authors mentioning optimization of G-theory designs, creation of standard errors (and possible confidence bands) for individual observed scores, and introduction of auxiliary information within G-theory designs among other things. However, the manuscript, in its present form, offers very limited information that would aid medical practitioners. In general, the information is not technical enough in some places and too technical in others. I realize that bridging the gap can be challenging and have suggestions below that might help. The manuscript also needs some serious editing and reorganization that is beyond the scope of what a reviewer should provide. Below I share some general comments that might improve the manuscript.

1. I would retitle the paper to describe what it is about more accurately. Overall, I presume the intent is to summarize recent developments in G-theory that might be of interest to medical educators. One suggestion for a title might be “Recent Advances in Generalizability Theory to Improve Assessments in Medical Education and Practice.”

**Response:** We have revised the title to “Recent Advances in Generalizability Theory to Improve Assessments in Medical Education and Practice”, which indeed matches the content better than the previous one.

2. I would start the paper with a general description of why reliability estimation is important, what traditional applications of univariate G-theory uniquely provide beyond conventional indices and then comment about multivariate G-theory. Two key benefits of multivariate G-theory are to provide better estimates of score consistency for composite scores and to correct correlation coefficients among measures for measurement error. I know that these are touched upon a bit in the current draft but can be better explained. I also would not make validity a prominent part of the paper or even mention it explicitly. You seemingly are referring to the disattenuated correlations alone, which are only one example of internal evidence related to validity.

**Response:** The beginning section has been revised according to the Reviewers’ suggestion: it emphasizes the importance of reliability and how G-theory relates to reliability estimation. Also, we have added “…Using multivariate G-theory can yield better estimates of score consistency for composite scores and correctly estimate correlation coefficients among subdomains for measurement errors when corresponding G-theory design demands assumptions for the existence of multiple subdomains….”

In addition, we appreciate the caution in mentioning validity, and therefore, the statement about “validity” has been wiped out in most parts of the manuscript. However, in the conclusion section, we have touched discussion about the validity of G-theory in the paper, especially the emphasis on using it with Kane’s framework, which begins with “While G-theory is recognized as a method for estimating the precision of measures, its uniqueness in partitioning and quantifying variance can also yield validity evidence….”

3. When describing the material above and elsewhere, you should provide concrete examples of how that information would be used to advantage in medicine. When possible, describe specific types of measures used in practice. This will aid comprehension for readers with limited backgrounds in measurement and statistics which is seemingly your intended audience. For each advance, direct readers to program code and other resources that would help them in applying the procedures.

**Response:** We have added a walkthrough example consisting most the G-theory advances mentioned in this manuscript, so that readers can better follow the idea.

4. After the general introduction, I would individually highlight the recent developments in a logical order under appropriate headings. A table listing these developments along with relevant citations would help to direct readers to places where they can find further information. Of particular interest would be supplements or appendices with computer code within published articles that illustrate how to apply the techniques directly. I know Vispoel and colleagues routinely provide such things. Jorgensen (2021) and others cited to do so as well.

**Response:** We really like the idea of having a table to summarize the content. Table 1 has been added to better present the points.

5. There seems to be some confusion regarding dimensionality. Using univariate G-theory does not mean that scores are unidimensional. It only means that a single measure was used. Similarly, multivariate G-theory simply means that multiple measures were used. Scores from each measure need not be unidimensional or even collectively multidimensional. G-theory analyses make no assumptions about score dimensionality (see Cronbach et al., 1963). In addition, poor model fit indices would not invalidate a G-theory model.

**Response:** We were meant to mention multiple dependent variables, instead of dimensionality like IRT assumes. We thank the Reviewer for discovering the confusion. The terms have been deleted.

6. The number of footnotes seems excessive in such a short article. Try to make the points clearer in the text itself.

**Response:** We have substantially extended the manuscript to make it a long article. But we have also paid more efforts to make the points clearer by providing details and examples.

7. Techniques for assessing subscale added value within multivariate designs needs to be better explained and expanded if included.

**Response:** We have extended the subscore literature in the manuscript to better describe subscales’ added value: “…It’s intuitive that subscores can provide potential diagnostic value for examinees interested in knowing their strengths and weaknesses in specific content areas,.. if the indexes aren't sufficiently high (e.g.,0.8), it will be likely to hint that the subscore result is misleading, and therefore should not be reported…”

8. You make contradictory statements about the extent to which G-theory has been applied in the research literature. Sometimes you say that applications are plentiful and at other times scarce. I personally think that applications of multivariate G-theory in the research literature (and field of medicine in particular) are very limited. This is not surprising given that sufficiently large samples would be needed when applying G-theory to performance assessments in medical settings. However, in the meta-analysis you cite, there seemingly are a larger number of relevant applications. All of this should be clearly explained. To enhance overall clarity further, you might come up with an ongoing example, even if hypothetical, that could be applied throughout the paper. You could then explain how each new development could improve assessment for your illustrated example.

**Response:** We appreciate this suggestion as it indeed sounded very contradictory. In the revision, we have made a clearer statement that “…even though the volume of publications related to multivariate G-theory is sufficient for meta-analyses, it doesn’t alter the fact that, compared with univariate G-theory or classical test theory, the applications of multivariate G-theory in medical educational research are relatively small…”and “…Also, samples required for multivariate G-theory would be even larger, making many applied studies not qualified for using the framework…”

9. Some recent work regarding advances in G-theory are seemingly omitted, especially that of Vispoel, Jorgensen, and perhaps others. For example, Vispoel and colleagues have applied structural equation modeling to univariate, multivariate, and bifactor G-theory designs to reproduce complete variance and covariance components; represent congeneric relationships; adjust for binary and ordinal data; and produce confidence intervals for key parameters. Some of their latest work may appear in conference papers, in press articles, and submitted articles under review. To keep the present paper current, I would recommend that you contact these researchers for such information.

**Response**: We have recognized Vispoel, Jorgensen, and other researchers’ works related to G-theory development; they have made a great contribution to the literature. For example, three references in our previous submission attribute to Vispoel and his colleagues.

To keep the present paper updated, we have added more recent works from these researchers (e.g.,Vispoel, Jorgensen, and others) with proper discussions and extensions. We have also endorsed these contributions at the part of “Generalizability Theory”section as evidence of the need for advancing G-theory in assessment in medical education.

Reviewer: 2

Comments to the Author

The paper aims to introduce several under-used features of G-theory to a general audience in medical education. The authors acknowledge the vast research applying G-theory to assess reliability in medical education, making this paper a welcome addition. While the paper is well-written and engaging, I have some suggestions for improvement.

First, I recommend adding subsections within the main body to improve the paper's structure. This will make it easier for readers to identify the core content of each paragraph and section. For example, the authors can create subsections for conditional standard errors of measurement, cut-score specific coefficients, etc. Additionally, I noticed that the discussion on the joint use of SEM and G-theory and multivariate G-theory wasn't highlighted earlier in the section (page 6). I suggest the authors address this issue.

**Response:**

Second, the paper lacks a thorough review or summary of the relevant research. As the study does not include empirical data analyses, the authors can contribute to the field by summarizing the research conducted using G-theory in medical education. I recommend adding a table that groups studies according to the topics discussed in the paper, highlighting the features utilized in each study and what is missing.

**Response:** We really like the idea of having a table to summarize the content. Table 1 has been added to better present the points.

Third, some of the sentences in the paper are misleading and require revision. For example, on page 7, the authors stated that reporting Cis may enrich the interpretation of results. However, this practice is rather common in the literature. Even a cursory search of google scholar using the terms “performance assessment” “medical” “generalizability theory” and “confidence interval” yielded 151 results. On page 8, the authors suggested that simply summing the estimates is frequently adopted, while G-theory mostly utilizes the mean-score metric. This sentence could potentially mislead readers. Also, concerning reporting a cut-score specific coefficient, I am not quite sure if there is such methodology in G-theory (I personally have not heard of but if there is, please provide references).

**Response:**

Fourth, some of the sentences require further clarification. For example, on page 5, the authors stated that the discrepancy of the dimensionality assumption does not alter the properties of G- and D-studies. I suggest the authors explain this statement further. On page 10, the authors stated that the solutions tend to be subjective. However, the solution found is often the one yielding the highest reliability. I recommend that the authors clarify why the solution is subjective.

**Response:**

Finally, on page 12, the authors stated that G-theory can be converted to SEM. I suggest the authors rephrase this sentence to avoid misleading readers. G-theory and SEM are distinct entities and cannot be converted from one to another. However, analytical solutions can be found using the same estimation tool. Also, in the Abstract, “ensuring that the assessment measures what it …” this is the definition of validity while G-theory mostly addresses reliability, not validity.

**Response**: We have deleted the inappropriate expression and re-written the sentences according to the Reviewer’s suggestion: “…It should be noted that, the vein of SEM for G-theory estimation does not imply that they can be converted….”We have also changed the abstract part to“…to navigate the investigation on that the reliability of assessment measures…,”to respond to the Reviewer’s comment.

Reviewer: 3

Comments to the Author

Thank you for allowing me the opportunity to review the manuscript by Ouyang et al titled ‘Advanced the Application of Generalizability Theory to Performance Assessments in Medical Education.’

While I appreciate the authors’ thorough review of ‘less-known, yet advanced and valuable, G-theory properties’ (page 2, line 30), I do not believe they met their objective of allowing the reader to advance their application and/or utilization of G-theory, specifically to performance assessments in medical education. The authors identify in their conclusion that this manuscript is ‘meant to be non-technical’ and ‘aims at informing medical education researchers about what G-theory can provide more to performance assessments.’ (page 14, line 8). However, after reading the paper multiple times, I found it to be extremely technical in the sense that the intended audience is likely statistical analyzers of medical education research. I am a clinical educator as well as a medical educator leader, and therefore I might not be the target audience for this manuscript and discussion. But, I believe myself and my colleagues are the intended audience for the Annals of Medicine Medical Education Section.

**Response:** Thank you for pointing drawbacks of this paper, and we realized that, although the paper was meant to be less-technical, the reading was hard to follow for an educator and/or applied researcher. Therefore, to increase the readability and practicality of this paper, we have revised/added two components:

1) changing the title to “Recent Advances in Generalizability Theory to Improve Assessments in Medical Education and Practice”

2) adding an example before the conclusion section to let non-technical readers better follow the content.

However, we do like to emphasize again that, after knowing the “advances” part in the paper, researchers may still need a statistician/technician to execute the high-level functions.

I have some concerns/recommendations for the authors:

Abstract:

-Page 2, line 40. I do not believe it is common to note in the abstract referring to the introduction of the paper and specifically indicating your goal to outline your manuscript ‘as the most minor technical as possible.’ Instead, the abstract should be a traditional abstract; describing your paper in an abbreviated format.

**Response:** The goal has been deleted.

-Page 2, line 32: I do understand what it means to ‘spin control the status quo’. From the context, I understand your point, but I’m not familiar with that term.

**Response:** The term has been removed.

Introduction:

-I thought the introduction was organized and outlined your goals, specifically point #4 on the scientific procedures or data analysis and interpretation that you address in the rest of the paper. I was also interested in finding out how your manuscript would help outline G-theory application in low-stakes settings.

**Response:** We have weakened the emphasis on high-stakes settings, as most properties are also appropriate for low-stakes settings. And we have added “…A typical instance is a mini-clinical evaluation exercise (mini-CEX) for residency training….”However, we have highlighted the part about cut-score to be more specific for high-stakes environments.

Generalizability Theory:

-Page 4, line 50. It’s also uncommon to start a section with a question and I would recommend that your reward your first line into a statement.

**Response:** The questionhas been turned into a statement.

-Page 6, line 25. Again, I was interested to read more with your goal to ‘inform researchers what could have been achieved with more G-theory.’

**Response:** Hopefully the added components and the revised parts highlighted in this version can make this statement more convincing.

Advancements of G-theory Application:

-I believe this entire section needs to be reorganized and clarified. It is essentially 8 pages of what I believe are the ‘lesser known G-Theory properties. However, after reading this section, I’m unclear what those properties are and how they can be applied to ‘Performance Assessments in Medical Education’ and specifically low-stakes assessments that medical educators frequently encounter.

**Response:** A As mentioned above, we have weakened the emphasis on high-stakes settings. Psychometric properties and statistical decisions are essential for evaluating the assessment in both low-stakes and high-stakes settings. Moreover, the cost-effectiveness analysis within G-theory framework is mentioned in the manuscript; It can be beneficial for OSCEs planning because even low-stakes OSCEs can be costly.

-I would suggest breaking this section into sub-categories and highlighting and defining or even describing these ‘lesser-known G-theory properties.’ I think the groups would include multivariant G-theory, D-studies (?), structural equation modeling, then maybe the three programs used (G\_String, GENOVA Suite and EduG.). My overall point is that I do not understand the basics of any of these and your manuscript jumps right into the heart of the literature review instead of framing their background and showing the reader how to use these concepts in their future G-theory performance assessments.

**Response:** In addition to revisions listed above, such as adding a summary table and a walkthrough example, we have adopted the sub-categories as the Reviewer suggested. Hopefully, with all new efforts, the revision can make more sense.

Reviewer: 4

Comments to the Author

1. Page Number 2: Key message heading: Key message needs to be precise / to the point and single line.

**Response:** We have re-adjusted the key messages.

1. The overall article does not seem to be a commentary type. It seems a review kind of paper.

**Response:** We have considered all reviewers’ comments to make the writing more suitable for the genre.

3. Page number 5, line number 4: check to spell non-negligible.

4. Page number 5, line number 26. D-study needs to be defined in full form.

**Response:** We have added decision study as the definition of D-study.

5. Overall document: Reference style must be uniform according to the journal.

**Response:** We have revised the references to make them consistent with the requirement.

6. In the conclusion part, page 13: References mentioned in the conclusion are not required. This needs to be checked according to the journal policy.

**Response:** We have deleted the references from the conclusion section.