A Comparison between Qualitative and Quantitative Approaches to the Fairness Issues in Educational Assessment

Tony C. A. Tan

Centre for Educational Measurement, University of Oslo

UV9030 Research Design

Prof Øistein Anmarkrud & Prof Marte Blikstad-Balas

30 October 2021

A Comparison between Qualitative and Quantitative Approaches to the Fairness Issues in Educational Assessment

This paper compares the research designs of the following two papers: Korobko et al. (2008) and Tierney (2014). Both articles can be obtained by following the DOI hyperlinks in the reference list.

Background and Rationale

Fairness is a central concern for all educational assessment. Since test results are often used for distributive purposes with high-stakes consequences, it is imperative for educators to ensure the assessment products they compiled are "fair". It was not until the 2014 publication of *Standards for educational and psychological testing* (the *Standards*, AERA et al., 2014) that a definition of testing fairness was explicitly stipulated:

> All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population. (p. 63)

The two major branches of educational assessment are formative and summative assessment, with the former serves the function of assessment *for* learning and the latter being the assessment *of* learning. This assignment therefore selected one publication for each category with a qualitative study by Tierney (2014) addressing classroom assessment fairness and a quantitative paper by Korobko et al. (2008) on the scoring procedures of grade point averages (GPA) used in the Netherlands' 1994/1995 academic year.

Overview of the Two Papers

The two papers selected for this assignment were both drafted prior to the publication of the *Standards* (2014) but their concept constructions of assessment fairness were highly consistent with the *Standards*. Korobko et al. (2008) operationalised assessment fairness as the comparability of subject difficulties in computing the GPA scores while Tierney (2014) positioned her enquiry at the classroom assessment level. In addition, Korobko et al. (2008) represented the mathematical

statistics end of the quantitative–qualitative spectrum in conducting educational research, while Tierney's (2014) case study design occupies the opposite end.

Purpose Statement

A purpose statement declares the intent of the entire paper by explaining why this study was conducted and what it intends to accomplish (Creswell & Creswell, 2018). The Tierney (2014) paper used signpost words such as "purpose" to help readers locate its central controlling idea. It also narrowed the study onto a single phenomenon of classroom assessment and used an action verb "illuminate" to convey how research learning will take place. The author justified the multi-case study design as her strategy of inquiry, as well as her criteria for including the particular six participants in her final report. The site (high schools in Ontario, Canada) and scope (at least 10 years of teaching experience, held a relevant undergraduate degree, and specialists in teaching English Language Arts) of the participation were also clearly reported.

The purpose statement in Korobko et al. (2008) was more implicit. Although signpost words such as "purpose" and "objective" were present in this paper, they were not used for articulating the purpose statement. The authors did, however, spent large portion of their publication on identifying the theory, model and conceptual framework, such as item response theory (IRT). Due to its advanced mathematical nature, this paper did not state explicitly the independent and dependent variables or to put them in an order in accordance with the purpose statement. The "An Example" section of Korobko et al. (2008, pp. 147–153) applied the statistical model derived earlier to an archival dataset collected by the Dutch Inspection of Education, implicitly stating this study's strategy of inquiry. Both the number of items ("60 fairly common combinations of examination subjects", p. 147) and participants ("The resulting data set consisted of the examination results of 6,142 students." p. 147) were clearly reported. Korobko and colleagues (2008) provided minimum definitions of key variables probably under the assumption that their readers were already well acquainted with these terminologies from prior training. Overall, the purpose statement of Korobko et al. (2008) appeared scattered and implied.

The two selected papers declared their purpose statements in drastically different styles. While Tierney (2014) closely followed the best practice checklist (Creswell & Creswell, 2018), Korobko et al.'s (2008) quantitative paper focused more on

mathematical rigour than on text formality. Tierney (2014) walked with her readers through the complex insight of classroom assessment whereas Korobko et al. (2008) assumed substantial prior knowledge from their readers in order to appreciate their goals and contributions.

Methodology and Analytic Approaches

Quantitative and qualitative research methods differ substantially between the two papers and shall be discussed separately. Korobko et al.'s (2008) main focus was statistical method *development*, with a data analysis section for verification purposes. It therefore did not fit neatly into Creswell and Creswell's (2018) description of quantitative methods (Chapter 8) whose main audience would expect statistical method *applications*. The appropriateness of Korobko et al.'s (2008) method-building process is accessible mostly to "insiders" of the quantitative research circle while appearing abrupt and poorly motivated from an outsider point of view. Korobko et al. (2008) indeed followed the tradition of the incremental model-building process widely shared in the quantitative community by starting with a simple model under the naïve assumption of unidimensionality (Model 1, pp. 148–150), then relax this constraint to accommodate multidimensionality (Model 2, pp. 150–151), and lastly let students' subject choice behaviour interact with this multidimentional model (Model 3, pp. 152–153). The authors have constructed their modelling architecture using generalised partial credit model (GPCM, Muraki, 1992) and have chosen marginal maximum likelihood (MML, Bock & Aitkin, 1981) as their estimation procedure with appropriate attribution to the corresponding landmark papers. It is not uncommon for quantitative papers to refer the readers to original publications for motivation and technical details so that more real estate can be devoted to modelling improvement. When a less well-known procedure was introduced such as the comparable fit statistic in the "Model Fit" section (pp. 145–146), the authors carefully laid out the background, the motivation, major citations and a brief mathematical derivation before applying such procedure to the data set. On balance, Korobko et al.'s (2008) methodology followed a well established series of steps. They presented their modelling in a logical order and immediately verified the applicability of these formulæ using a real-world data set before drawing their conclusion that the existing practice can be improved in certain way. This approach enhanced public's confidence in this paper's scientific rigour and the credibility of their

recommendations.

Tierney's (2014) qualitative research methodology, on the other hand, followed closely Creswell and Creswell's (2018) prescription (Chapter 9, pp. 179–211). The author started with a review of potential audiance's needs by pointing out that there lacked consensus on classroom assessment fairness definition and operationalisation. The nuanced debate about many interpretations of fairness in assessment theory was then presented, differentiating fairness from (statistical) unbiasedness and the concept of test validity. This paper then resented the "researcher as key instrument" (Creswell & Creswell, 2018, p. 181) by collecting interview data herself through interviewing participants using a protocol involving vignettes and interpret subsequent responses into multiple facets. Tierney paid particular attention on justifying the appropriateness of her chosen strategies for establishing qualitative research rigour. She then purposefully selected participants for her subsequent analyses and reported the recruitment strategies, the reasons behind selecting these particular six participants and their attributes (Table 2, Tierney (2014), p. 58), as well as the types of data to be collected (both vignette responses and subsequent open-ended interviews). All this information enhanced readers' confidence in the credibility of data source, upon which common themes were extracted and validated. Tierney's (2014) methodology section is particularly transparent and well structured. It enabled verification by both insiders and outsiders of the qualitative research community. Such falsifiability promoted Tierney's (2014) paper from an opinion piece to a reliable scientific contribution, whose taxonomy of classroom assessment fairness can be relied upon by subsequent researchers.

Both papers were strong in their methodologies and analytical approaches. Although Korobko et al.'s (2008) structure fell outside of traditional textbook prescription, its incremental model-building approach combined with an immediate empirical verification served to establish this paper's scientific rigour for evaluating differences in GPA subject difficulties. Similarly, by following commonly accepted protocol in the qualitative research community, the Tierney (2014) paper stood firmly on its own as a trustworthy source of information for establishing teachers' interpretation of classroom assessment fairness issues.

Findings

Both papers provided unambiguous findings and recommendations. Using rigorous and sophisticated statistical techniques, Korobko et al. (2008) was able to demonstrate the inadequacy of unidimensional IRT models in favour of multidimensional constructions for assessing inter-subject difficulties. The optimal model was found to be one that explicitly accommodated students' subject choice behaviour, with adjusted GPA being the preferred policy recommendation over "raw" GPA scores. Tierney's (2014) qualitative derivation also put forward generalisation along two axes: transferability and theoretical generalisation, with the former meaning the generalisation into similar contexts while the latter helping nuance and update evolving theory. The author concluded her writing with two future research questions related to teacher education for the purpose of enhancing teachers', particularly novice teachers', assessment literacy. Findings from both papers carried significant weights in both methodology and practicality; they collectively formed a three-dimensional view on the educational assessment fairness issues, complementing each other in research traditions (quantitative vs qualitative) and assessment domains (summative vs formative).

Quality Evaluation

The two chosen papers approached the fairness issues in educational assessment from different epistemological foundations. Korobko et al. (2008) took a postpositivism worldview by reducing the educational assessment process into quantifiable parameters (the θs and βs in their equations) and by assuming a deterministic relationship between these parameters and students' GPA outcomes. It used empirical observations and measurement to verify a series of IRT models by incremental model improvement. Tierney (2014), on the other hand, took a constructivism worldview with a chief purpose of understanding the interpretations and meanings from six purposefully chosen participants. Through social construction, this inductive exercise generated a theory of classroom assessment fairness framework. Both papers fulfilled their research purposes via rigorous methodologies and careful analyses. The resulting findings were well grounded, trustworthy and actionable. It is therefore in this assignment's opinion that both Korobko et al. (2008) and Tierney (2014) were high quality publication with strong research designs.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* American Educational Research Association; American Psychological Association; National Council on Measurement in Education. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Creswell, J. W., & Creswell, J. D. (2018). *Research design* (5th ed.). SAGE.

Korobko, O. B., Glas, C. A. W., Bosker, R. J., & Luyten, J. W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement*, *45*(2), 139–157. https://doi.org/10.1111/j.1745-3984.2007.00057.x

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), 1–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Tierney, R. D. (2014). Fairness as a multifaceted quality in classroom assessment. *Studies in Educational Evaluation*, *43*, 55–69. https://doi.org/10.1016/j.stueduc.2013.12.003