



OPEN ACCESS

EDITED AND REVIEWED BY

Gavin T. L. Brown,
The University of Auckland,
New Zealand

*CORRESPONDENCE

Tine van Daal
tine.vandaal@uantwerpen.be

SPECIALTY SECTION

This article was submitted to
Assessment, Testing and Applied
Measurement,
a section of the journal
Frontiers in Education

RECEIVED 16 November 2022

ACCEPTED 28 November 2022

PUBLISHED 12 December 2022

CITATION

van Daal T, Lesterhuis M, De Maeyer S
and Bouwer R (2022) Editorial: Validity,
reliability and efficiency of
comparative judgement to assess
student work. *Front. Educ.* 7:1100095.
doi: 10.3389/feduc.2022.1100095

COPYRIGHT

© 2022 van Daal, Lesterhuis, De
Maeyer and Bouwer. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

Editorial: Validity, reliability and efficiency of comparative judgement to assess student work

Tine van Daal^{1*}, Marije Lesterhuis², Sven De Maeyer¹ and
Renske Bouwer³

¹University of Antwerp, Antwerp, Belgium, ²University Medical Center Utrecht, Utrecht, Netherlands,
³Utrecht University, Utrecht, Netherlands

KEYWORDS

comparative judgment (CJ), validity, reliability, efficiency, performance assessment

Editorial on the Research Topic

[Validity, reliability and efficiency of comparative judgement to assess student work](#)

Assessing complex skills such as writing, designing, or problem-solving is a challenge. Comparative judgement is considered to be a reliable and valid method for assessing student work (e.g., [Lesterhuis et al., 2018](#); [Verhavert et al., 2019](#)). In comparative judgement, students' work is evaluated by pairwise comparison. As assessors only must indicate which piece of work is better, differences in severity are not at play ([Pollitt, 2012](#)). Furthermore, each work is compared with several others and evaluated by multiple assessors. Based on these comparisons, the quality of each individual work can be estimated. This quality score reflects, so to speak, the shared consensus of the assessors ([Jones et al., 2015](#); [van Daal et al., 2019](#)).

The comparative judgement approach is based on Thurstone's law of comparative judgement (1927), which states that it is possible to discriminate between objects on a single scale through a series of pairwise comparisons ([Thurstone, 1927](#)). Even though Thurstone already proposed the possibility of using comparative judgement for assessment in education, it was not until 2004 that Pollitt introduced the method in education in his paper "Let's stop marking exams". His work convincingly explained the merits of comparative assessment in terms of validity and provided the first evidence for a reliable summative assessment. Now, almost two decades later, various comparative judgement tools are available for education, such as Comproved or NoMoreMarking. Moreover, researchers around the world have investigated the quality of the method, where and/or how it can be applied, and how the method can be improved.

In this Research Topic, we aim to provide a state-of-the-art of research on comparative judgement in education. We bring together current insights on the validity, reliability and efficiency of the method. In their contributions to this Research Topic, the

authors present recent empirical research, each with their own approach, perspective and research focus. In this way, this Research Topic offers the foundation for future research into comparative judgement.

How valid is comparative judgement?

Up to now, only a limited number of studies dig into the validity of comparative judgement (Whitehouse, 2012; Lesterhuis et al., 2018; van Daal et al., 2019) while this is crucial in light of the use of the scores resulting from comparative judgement (Messick, 1989). More studies into the validity of comparative judgement are highly needed to explore the validity of comparative judgement and factors that might affect the validity of the outcomes (Bejar, 2012).

An important line of research in this Research Topic is focused on the validity of comparative judgement to assess students' competences. Buckley et al. conducted a critical review of how ACJ (adaptive comparative judgement) has been used and studied in the field of technology education. They conclude that there is a need for more critical studies on the internal validity, a theoretical framework, and the consideration of falsifiability. Two studies in this Research Topic add to our knowledge base regarding construct validity, concurrent validity, convergent validity and predictive validity. Mentzer et al. first conducted a content analysis of students' work that was ranked high and low based on a peer assessment making use of CJ, concluding that there is evidence for construct validity. Then they examine the relation between scores obtained through peer assessment, instructors' assessment and students' final grades, concluding that students' peer assessment is an indicator of their final grades (predictive validity) but not an indicator of instructor scores (concurrent validity). Landrieu et al. investigated the extent to which comparative judgement scores converge with absolute analytic and holistic scoring methods. Results show that even though scores generated by the three methods highly correlate, there is substantial variation between methods in the information it gives to researchers and practitioners. This implies that one should consider the goal of an assessment when choosing one of the scoring methods. The authors conclude with an outline on the advantages and disadvantages of each of these methods.

In this Research Topic, there are two studies that investigated specifically the construct validity of comparative judgement. Chambers and Cunningham questioned whether assessors are affected by construct-irrelevant aspects of text quality when comparing texts in an experimental design. They conclude that judgements are influenced by handwriting and the presence of missing responses, showing that some biases might be at play when assessors compare texts. Lesterhuis et al. investigated whether assessors differ in how they evaluate

students' work using comparative judgement. More particularly, the authors examined to what extent we can distinguish between different types of assessors based on the aspects they take into account when comparing argumentative texts. Results show that assessors are comparable considering the aspects they evaluate during the process of comparative judgement, but that they are different in the weight they give to some aspects over others. This implies that for valid comparative judgement scores, it is warranted to include multiple assessors.

How reliable is comparative judgement?

Reliability is another important indicator of the quality of an assessment. Most of the early studies on comparative judgement focused on reliability, as it is especially high reliability in which comparative judgement stands out compared to other methods (Pollitt, 2012). In comparative judgement, the scale separation reliability (SSR) is used as indicator for reliability (Verhavert et al., 2018). Crompvoets et al., however, questioned this coefficient. They investigated the bias and stability of the SSR in relation to the number of comparisons per assessed work based on a simulation study. They conclude that the SSR can still be used as an indication of the reliability, even when the variance of the items is overestimated. However, they also recommend to obtain a sufficient number of comparisons per student work (i.e., 41 comparisons per item) to prevent an overestimation of the reliability by the SSR.

How efficient is comparative judgement? New applications, approaches and algorithms

As reliability and efficiency always seem to be a trade-off, it is not surprising that this Research Topic comprises a number of studies on ways to increase efficiency without compromising reliability and validity. Humphry and Bredemeyer show how different sets of works can be efficiently linked using a core set. Verhavert et al. also examined how new student works can be placed in an efficient and reliable manner on a previously calibrated reference set. They conclude that this alternative application of comparative judgement does not hamper the reliability of scores. Seery et al. outline how CJ can be used as a vehicle to set nation-wide standards and unravel teacher constructs of quality at the same time. Benton describes a simplified pairs approach to increase efficiency in the context of equating standards in high-stakes contexts. His simulation study underpinned its superior accuracy to current approaches. De Vrindt et al. investigated whether and how *text mining* can help to make a CJ assessment of textual products more efficient by taking into account information gained through the text

mining in the selection of pairs for CJ. They show that the use of this technique increases efficiency while reducing inflation of the reliability estimate used in CJ. Leech et al. approached CJ as a method for equating the standards set in high-stakes testing contexts, to assure that marks are comparable over the years. They investigate the link between the number and length of tasks and the difficulty of the comparison. They compared the outcomes with the outcome of another method—namely traditional equating—and ask assessors about the judgement processes. They conclude that judges used similar processes in CJ within a topic, but over topics there were differences in how judges come to a decision, making the authors discuss the ability of CJ to maintain an audit of how decisions are made.

Implications for educational practice and future research

This Research Topic shows that applications of comparative judgement widely differs in practice. First, different types of competences and student work are assessed (writing, chemistry, mathematics) demonstrating that the application of comparative judgement is not restricted to a single educational domain. Also, the contributions in this issue show that comparative judgement can be used for different purposes such as peer assessment, instructor assessment, standard setting, and equating. Finally, this Research Topic also demonstrates that the methodology used in research on comparative judgement ranges from qualitative research on assessors' judgement processes, over experimental research to simulation studies. As such, by studying the merits and disadvantages of comparative judgement, the conditions and contexts of

comparative judgment have become an interdisciplinary field of research in itself, as demonstrated in this Research Topic.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the editorial and approved it for publication.

Conflict of interest

ML and SD founded, next to their academic position, Comproved. Comproved makes it possible for teachers to use comparative judgement in their classroom. They have, however, no financial returns from this company.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Bejar, I. I. (2012). Rater cognition: implications for validity. *Educ. Meas. Issues Pract.* 31, 2–9. doi: 10.1111/j.1745-3992.2012.00238.x
- Jones, I., Swan, M., and Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *Int. J. Sci. Math. Educ.* 13, 151–177. doi: 10.1007/s10763-013-9497-6
- Lesterhuis, M., Daal, T., van, Gasse, R. V., Coertjens, L., Donche, V., and Maeyer, S. D. (2018). When teachers compare argumentative texts: decisions informed by multiple complex aspects of text quality. *L1 Educ. Stud. Lang. Literat.* 18, 1–22. doi: 10.17239/L1ESLL-2018.18.01.02
- Messick, S. (1989). Meaning and values in test validation: the science and ethics of assessment. *Educ. Res.* 18, 5–11. doi: 10.3102/0013189X018002005
- Pollitt, A. (2012). The method of adaptive comparative judgement. *Assess. Educ. Principl. Pol. Pract.* 19, 281–300. doi: 10.1080/0969594X.2012.665354
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychol. Rev.* 34, 273–286. doi: 10.1037/h0070288
- van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., and De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assess. Educ. Principl. Pol. Pract.* 26, 59–74. doi: 10.1080/0969594X.2016.1253542
- Verhavert, S., Bouwer, R., Donche, V., and De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assess. Educ. Principl. Policy Pract.* 26, 541–562. doi: 10.1080/0969594X.2019.1602027
- Verhavert, S., De Maeyer, S., Donche, V., and Coertjens, L. (2018). Scale separation reliability: what does it mean in the context of comparative judgment? *Appl. Psychol. Meas.* 42, 428–445. doi: 10.1177/0146621617748321
- Whitehouse, C. (2012). *Testing the Validity of Judgements About Geography Essays Using the Adaptive Comparative Judgement Method*. Manchester: AQA Centre for Education Research and Policy.