Book review: Measuring writing

Van Steendam, E., Tillema, M., Rijlaarsdam, G., & Van den Bergh, H. (Eds.) (2012). Measuring Writing: Recent Insights into Theory, Methodology and Practices. [Studies in Writing. Leiden/Boston: Brill | ISBN: 978-1-78190-267-7

Reviewed by: Mariëlle Leijten, Research Foundation – Flanders / University of Antwerp

Establishing text quality is often an important aspect of developing understanding both of writing processes and writing interventions. However, this concept of 'text quality' is often not very well operationalized or described. The editors of this book state its aim as being to facilitate developments of a more precise understanding and definition of writing ability. The book comprises eight chapters that can be roughly divided into these topics: definitions of writing ability, rating procedures, rater effects and automated essay scoring.

Definitions of writing ability

In the first part of the book two chapters deal with the question: Is writing ability generalizable across contents? And, a related question, how many texts must be sampled from a student before something can be said about their general writing ability. In the chapter by Schoonen, he illustrates theoretical considerations about validity and generalizability by analyzing data from a large group of students at three moments in time, for three writing tasks in two languages. Schoonen states that for secondary school students 3 to 4 assignments, each rated by two raters, will be sufficient to make a general claim about a student's written English as a Foreign Language ability. However, in L1 7- to 10 double-rated assignments would be necessary.

journal of
WRITING RESEARCH

Leijten, M. (2013). Measuring writing [Review of the book *Measuring Writing: Recent Insights into Theory, Methodology and Practices* by E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. Van den Bergh, H. (Eds.)]. *Journal of Writing Research, 6*(1), 85-88. http://dx.doi.org/10.17239/jowr-2014.06.01.4

Contact: Mariëlle Leijten, University of Antwerp, Department of Management, Prinsstraat 13, 2000 Antwerpen | Belgium - marielle.leijten@uantwerpen.be

So, unfortunately, results show that generalizability is quite limited in L1 tasks, since intra-writer variability is larger than in L2-writing. The results also show that adding tasks to the assessment is more beneficial for the generalizability than adding raters. Also the study by Van den Bergh et al. shows that writing in a foreign language is a more stable activity than L1 writing, especially in older students (1st year university students vs. 9th graders). Holistic scoring via benchmarking with typical examples and analytical rating on four dimensions (content, argumentation, conclusion and text structure) are compared. The analytic scores are reliable but quite topic-dependent, whereas holistic ratings are less reliable, but they are more generalizable. When striked by the high number of assignments advised by Schoonen, Van den Bergh et al. are a bit more moderate: "The bottom line of this study is that writing research with only one text as a criterion does not provide anything. […] studies in which only one text is used as a criterion cannot produce any reliable result." (p. 32). The same holds for number of raters. Multiple raters need to agree on the quality of preferably multiple texts.

Rating procedures

In this part three chapters discuss various rating procedures. Neumann uses different assessment studies to describe the advantages and disadvantages of various rating procedures: holistic, analytical and mixed models based on research results from the USA and Germany. Olinhouse, Santangelo and Wilson continue to describe the American standard of assessing writing. In America most tests are single-occasion, single-genre and holistically scored. However, these tests are used to interpret the result as level of proficiency of the totality of writing standards (which comprises more genres and skills). But, as the preceding chapters show writing in L1 is not a stable activity at all. Therefore, the inferences made from a single-occasion and single-genre that is holistically scored may be limited to the genre under study. As such, the large-scale assessments are providing too narrow a view of students' writing ability. Solutions need to be found that allow assessment on a broader range of tasks.

Rater effects

Varying genre might lead to interpersonal variability across writing tasks. Raters might also be a source of variability (see also He et al. 2013). In the third part, two chapters deal with this issue. In the contribution by Barkaoui and Knouzi six raters scored two tasks on about 20 features that could be subdivided in more than 60 (!) individual variables. Their argument for the extremely thorough approach is that writing studies should look beyond the test scores. This approach is very interesting from a research perspective, but not very realistic or useful in educational settings. Weigle and Montee focus on contemporary writing assignments by taking textual borrowing into account. They integrate reading and writing since this reflects contemporary writing in real-world academical and professional settings. It was shown that the five raters did not agree on how source texts could be incorporated in the essays. The opinions were even quite

divergent, showing the importance to provide clear instructions and trainings to raters, especially when they have a different background.

Automated essay scoring

The book ends with the solution that may by this point have already occurred to readers. Would it not be possible to automatize certain measures? This would make at least the rating part a bit more feasible. And, would it not be possible that some 'easy to define' measures strongly correlate with 'complex' measures so that you might settle with the first measure. In two chapters tools and techniques on automated essay scoring are described. McCurry provides an overview of IntelliMetric and E-rater (both American) computer scoring software of the Educational Testing Service (ETS). McCurry is quite doubtful in his review that computers can provide the same scoring validity as humans. Finally, Withaus discusses the challenges of Automated Essay Evaluation for evaluating multimodal writing. In addition to the tools discussed previously in the chapter by McCurry, Withaus focusses (as do Weigle and Montee) on contemporary multimodal writing. No solutions on this topic yet, just opportunities for future research agendas (also Burstein, Tetreault, & Chodorow, 2013; Deane, 2013). For more information on Automated Essay Evaluation, I refer to the Handbook of AEE: Current applications and new directions by Chermis & Burstein, 2013.

My opinion

I must admit that my goal in choosing to review this book was very much driven by wanting to know which measures I could best use in a study on multilingual essay writing processes. I know now that I should be using multiple-occasion, multiple-genre tasks, double-reviewed and assessed both holistically and analytically. But, was my initial question answered by reading the book? Or, can I point a new PhD-student to a chapter that provides the answer? I must answer these questions by saying 'no'. The book is very interesting if you want to know fine-grained information about aspects on scoring methods (including automated methods), procedures, and rater effects. The topics in the book are also well tied together, first, by the insightful and closely-argued introduction, and second by the individual authors. Numerous cross-references are made, contrary results are discussed and sometimes explained. I feel, though, that this level of details is also a shortcoming of the book. I, and I assume other researchers too, would like to be able to choose an assessment method that suits our research purposes: I would like to be able to choose a reasonable amount of texts, raters, etc.

After reading the book, you will probably be left with the feeling: I cannot accomplish this in this lifetime. I would need the participants to write at least 3 to 4 texts in their foreign language and 7 to 10 (!) in their native language, I would have the texts rated by a minimum of 4 raters, on at least 40 variables. So, research in the field of defining text quality still has important work to do in describing a method to conduct reliable and feasible research into the relation between writing processes and text

quality. The studies described are more fundamental and methdologicalstudies and might as such not be transferable to other more practically-oriented research projects.

But I think that the book makes it clear that it is not very sensible to have a small group of young students write a single text in their mother tongue and have this rated by one rater and then to expect this rating to provide a robust indication of their writing competence. As the introduction to the book indicates "[…] writing ability cannot be reliably assessed by means of a single writing product per writer. Multiple texts per writer and multiple raters per writing product are required. Additionally, raters are preferably trained, scoring rubrics need to be clear and detailed and preferably accompanied by benchmarks." (p. 10). Intra-writer task variability not only plays a role in text quality variation, but also in writing processes: one needs at least three processes (within a genre) in L2 and four in L1 to get an accurate indication of a students' typical writing process (see also Rijlaarsdam et al., 2011).

The problem raised in the introduction that writing quality measures are often not sufficiently well described to facilitate comparisons between studies might be partly caused by restrictions imposed by some journals. A restriction in length of the article might cause researchers to be very concise in describing measures. However, research in the field would obviously benefit from very thorough descriptions – as we have read in this book – and from sharing questionnaires and statistical techniques. Fortunately, other journals have become aware of this added value and require thorough material descriptions added to articles. Other journals offer to the possibility of making additional materials available online. This is the case with Journal of Writing Research.

Concluding, I wish for future researchers to continue developing suitable language dependent and language independent measures that, importantly, can be automatized and made freely available for the research community.

References

Burstein, J., Tetreault, J., & Chodorow, M. (2013). Holistic Discourse Coherence Annotation for Noisy Essay Writing. *Dialogue & Discourse, 4*(2), 34-52.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. Assessing Writing, 18(1), 7-24.

He, Tung-hsien, Wen Johnny Gou, Ya-chen Chien, I. Shan Jenny Chen, and Shan-mao Chang. 2013. Multi-Faceted Rasch Measurement and Bias Patterns in EFL Writing Performance Assessment. *Psychological Reports 112*(2):469-485. doi: 10.2466/03.11.PR0.112.2.469-485.

Shermis, M.D., & Burstein, J. (Eds) (2013). Handbook of Automated Essay Evaluation: Current applications and new directions. [Book review by Marie Stevenson]. *Journal of Writing Research, 5*(2), 239-243.

Rijlaarsdam, G., Van den Bergh, H., Couzijn, M., Janssen, T., Braaksma, M., Tillema, M., Van Steendam, E., Raedts, M. (2010). Writing. In S. Graham, A. Bus, S. Major, & L. Swanson (Eds.). *Application of Educational Psychology to Learning and Teaching. APA Handbook.* Volume 3 (p. 189-228). Washington, DC, US: American Psychological Association.