Chapter 2

Generalizability of Text Quality Scores

Huub van den Bergh, Sven De Maeyer, Daphne van Weijen and Marion Tillema

Abstract

This study shows that in order to measure writing skills reliably one needs multiple assignments rated by multiple raters. However, the number of assignments needed to assure a reliable measurement depends on the population studied. For first-year university students, one does not need to rate as many texts as for ninth-graders. Especially, when writing in a foreign language, first-year students prove to be very stable writers. The stability of text quality scores appears to depend on the way texts are rated as well. So-called analytic scoring schemes seem to result in reliable but topic-dependent text quality scores. Holistic ratings, on the other hand, appear to give raters less support (resulting in a lower inter-rater agreement) but result in less topic-dependent scores. Hence, in order to generalize, writers do not need to write that many texts if these texts are holistically rated. Implications of these results for experimental studies on writing and educational effectiveness of writing pedagogies are discussed.

Introduction

The quality of texts is rated in many studies on writing. Based on these ratings, inferences concerning students’ writing skills are made. That is, students with high text quality ratings are assumed to be good writers, whereas students with low text quality scores are assumed to write less well. In turn, such inferences are the foundation for conclusions on effectiveness of experimental educational programs.
(e.g. Braaksma, Rijlaarsdam, Van den Bergh, & Hout Wolters, 2004), or on relations between writing processes and writing skills (e.g. Van Weijen, Van den Bergh, Rijlaarsdam, & Sanders, 2009), and so on.

However, it is well known that different raters assign different quality scores to the same text. Hence, based on a single rating by one rater, one cannot draw firm conclusions on differences in quality between different texts (McColly, 1970; Nystrand, Cohen, & Dowling, 1993; Van den Bergh & Eiting, 1989). Therefore, in many studies inferences on writers’ writing skills are made based on jury ratings; the average of multiple individual raters is used as a measure of text quality. Of course, such a procedure is only warranted if there is (enough) agreement between ratings of different jury members. If the correspondence between raters is low, it does not make much sense to sum them to one jury rating, because (at best) they are all indicative of different aspects of text quality. In order to assess the correspondence between raters usually some measure of homogeneity is used. The most well-known measure is perhaps Cronbach’s alpha. More generally, the observed differences between observations ($S_{\text{observed}}^2$) can be partitioned into three types of variance: the variance between writers ($S_{\text{writer}}^2$), the variance between raters ($S_{\text{rater}}^2$) and the variance of the interaction between writers and raters ($S_{\text{writer}\times\text{rater}}^2$). Of course, the sum of these variance components equals the observed variance:

$$S_{\text{observed}}^2 = S_{\text{writer}}^2 + S_{\text{rater}}^2 + S_{\text{writer}\times\text{rater}}^2 \quad (2.1)$$

We are primarily interested in the variance between writers ($S_{\text{writer}}^2$), as this is an indication of differences in text quality between different writers. The second ($S_{\text{rater}}^2$) and third term ($S_{\text{writer}\times\text{rater}}^2$) can be considered as error terms. $S_{\text{rater}}^2$ denotes differences in leniency of individual raters, that is, the degree in which a rater is likely to give higher or lower marks in general, whereas the interaction ($S_{\text{writer}\times\text{rater}}^2$) is usually considered to be random noise. Based on estimates of these three variance components, the reliability ($\rho$) of the jury ratings can be approximated:

$$\rho = \frac{S_{\text{writer}}^2}{S_{\text{writer}}^2 + (S_{\text{rater}}^2/N_{\text{rater}}) + (S_{\text{writer}\times\text{rater}}^2/N_{\text{rater}} \times N_{\text{writer}})} \quad (2.2)$$

Text quality ratings are at best interval-scaled. As a result, the variance between raters can be neglected, and Formula (2.2) reduces to the well-known coefficient alpha (see, e.g., Osburn, 2000). So, based on the variance components related to different sources of variation the reliability of ratings of writing, or more general performance assessments, can be estimated.

Another but related issue concerns writers’ variability. It is well known that text quality, like other kinds of performance ratings, is likely to vary over assignments (or writing tasks). A writer can write a very good text for one assignment but a poor text for another assignment, whereas for another writer it is just the other way around. In the past, several explanations have been put forward for such differences in quality between different texts of the same writer. These explanations differ from referring to