Chapter 1

The Validity and Generalizability of Writing Scores: The Effect of Rater, Task and Language

Rob Schoonen

Abstract

In this chapter the validity and generalizability of writing proficiency scores are addressed within the framework of generalizability theory. Generalizability theory as a psychometric theory (Brennan, 2001) is a very suitable approach to deal with the multifacetedness of writing assessments and the different sources of ‘measurement error’ in such assessments (including raters and topic/task facets). At the same time, questions about the generalizability of scores raise related questions about the validity of writing assessments and about the construct of writing proficiency: What is the ‘universe of admissible observations’ for writing proficiency and what is the ‘universe of generalization’ to which we want to generalize? Generalizability connects the two main psychometric parameters, reliability and validity, as the two extremes of one and the same continuum.

The theoretical considerations about validity and generalizability are illustrated by analyses of empirical data. Writing scores of secondary school students are analysed from a generalizability perspective. Students responded to writing assignments in their dominant language (Dutch) and in English as a foreign language (EFL). Tasks were administered at three points in time and three tasks per language per measurement wave. Each task was rated by two raters (Schoonen et al., 2011). Generalizability of the writing scores is determined by taking into consideration effects of task and rater. Furthermore, generalizability of scores for the dominant language is compared to the generalizability of scores for EFL writing. It is hypothesized that the individual differences in FL writing
scores are more strongly related to differences in (language) proficiency, whereas scores in L1 writing are more affected by topical knowledge or other task effects. Analogously, increasing FL proficiency gives way to more influence of topic/task effects, which implies that the generalizability of the EFL writing scores is expected to decrease over time.

The results of the analyses show that the first hypothesis is corroborated by the data, but the second hypothesis is not. The outcomes are discussed within the framework of writing assessment in educational and applied linguistic research, but also within a broader context of validity theory.

Introduction

We can safely assume that proficient writing is the orchestration of many (psycho)linguistic and (other) cognitive skills that leads to the production of readable texts (see among others Grabe & Kaplan, 1996; Hayes, 1996; Schoonen, Snellings, Stevenson, & Van Gelderen, 2009). Analytic scoring or detailed analyses of the written texts could provide diagnostic information on the linguistic resources available to the writer, especially in second or foreign language writing contexts, and specific writing tests (e.g. structuring tasks or revision tasks) might give information on (some of) the writing processes involved in writing (Grabe & Kaplan, 1996; Schoonen, 2011; Weigle, 2002). However, often researchers and educationalists want an authentic or direct proof of a candidate’s overall proficiency, that is the candidate has to perform an actual writing task that is supposed to be close to real-life tasks, allowing a performance assessment. Because of its authenticity, we are inclined to accept the writing score derived from the writing performance at face value. Messick (1994) notes that performance assessments come in very different forms, and seldom we are interested in the performance, but more so in the product. This also goes for writing assessments in which the writing process (performance) is not at stake, but the piece of text that is the result of the writing performance. Furthermore, assessors might differ in their conceptualization of the assessment and the role of the writing tasks, writing tasks either being the target of the assessment and exemplar of other (real-life) tasks (task-driven) or being the vehicle to the assessment of certain skills and knowledge (construct-driven) (Messick, 1994). Especially in the latter case generalizability of scores is an important feature, that is, ‘... the meaning of the construct is tied to the range of tasks and situations that it generalizes and transfers to’ (Messick, 1994, p. 15). It is paramount that in the assessment all relevant

---

\(^1\)The use of the term ‘performance’ is to some extent ambiguous, referring to both the process of writing and the end product, that is, the text. Although process and product are related (cf. Breetvelt, Van den Bergh, & Rijlaarsdam, 1994), there certainly is no one-to-one relationship. The focus of this chapter is on scores for the written texts and with ‘writing performance’ we will primarily refer to the end product, that is, the text.