ELSEVIER

An investigation into current measures of linguistic accuracy in second language writing research

Charlene Polio [a,*], Mark C. Shea [b]

[a] Michigan State University, 251B Wells Hall, 619 Red Cedar Road, East Lansing, MI 48824, United States
[b] Mount Holyoke College, 50 College Street, South Hadley, MA 01075, United States

Abstract

This study begins by describing measures of linguistic accuracy in second language writing research. We first report on measures that have been used in the last ten years including holistic measures, error-free units, number of errors, number of specific error types, and measures that take error severity into account. We discuss differences in the measures used now vs. those reported on in Polio (1997) and conclude that interrater reliability and detailed coding guidelines are still underreported, making replication of studies very difficult. We then apply ten of these measures to the MSU data set to determine which are the most reliable and which best show change over time. The measures comprise holistic and error-free units as well as counts of specific error types. With the exception of counts of certain error types, we were able to obtain over .80 reliability on the measures though some of the measures were much easier to use than others. In addition, we attempted to determine whether or not these measures were measuring the same or different constructs by looking at correlations and change over time. Among the correlations, some were expected and some not. One conclusion was that weighted error-free units did not seem to differ from standard error-free units. Only the holistic measures and the number of preposition errors showed any change over time.
© 2014 Elsevier Inc. All rights reserved.

Introduction

Compared to other constructs discussed in this volume, accuracy is relatively easy to define. It has been defined as "freedom from error" (Foster & Skehan, 1996, p. 304) and "the ability to be free from errors while using language" (Wolfe-Quintero, Inagaki, & Kim, 1998, p. 33). Error, a more problematic term, has been defined as "a linguistic form or combination of forms which, in the same context and under similar conditions of production, would, in all likelihood, not be produced by the speakers' native speaker counterpart" (Lennon, 1991, p. 182). Dialect and register issues notwithstanding, there is some consensus among groups of language users as to what an error is.

Studies using measures of accuracy are frequent in the second language writing literature, and they have only increased with the explosion of studies examining the effects of written corrective feedback. To date, only Polio (1997) and Wolfe-Quintero et al. (1998) have provided comprehensive reviews of measures used to assess accuracy. Polio

---

* Corresponding author.
    E-mail addresses: polio@msu.edu (C. Polio), markshea@mtholyoke.edu (M.C. Shea).

(1997) reviewed studies that used some type of written accuracy measure as a variable and were published between 1984 and 1995 in seven applied linguistics journals. Wolfe-Quintero et al. (1998) examined the extent to which accuracy measures indicated development in 34 different studies conducted prior to 1996. They included three types of studies: longitudinal studies where accuracy was measured over time; cross-sectional studies where accuracy was measured among different proficiency groups; and correlational studies where accuracy measures were compared to holistic measures of writing proficiency or quality. Both of these studies are discussed in detail below.

Given the large number of studies that continue to claim to measure accuracy, the construct needs to be revisited with two purposes. First, the reliability of the recent measures must be examined in light of more recent calls for replicable research (e.g., Polio, 2012a; Porte, 2012; Porte & Richards, 2012). Second, we need to consider if and how accuracy develops over time and which measures might best capture development. The reliability of accuracy measures is important because the lack of reliable measures has contributed, in our opinion, to written error correction controversy: findings are not uniform across studies at least in part because of different and unreliable measures. An investigation into the relationship between accuracy and development will add to our ability to categorize the changes over time that comprise development and shed light on possible interactions between accuracy and other aspects of development, such as complexity.

Previous discussions of the reliability and validity of accuracy measures

Reliability

Four studies have explicitly addressed the issue of the reliability of measures of written accuracy. Wolfe-Quintero et al. (1998) found 34 studies that used a measure of written accuracy and they found that only seven studies reported either intra- or interrater reliability. Polio (1997) reviewed 16 studies that measured accuracy for a variety reasons, including to examine the effects of feedback (e.g., Kepner, 1991), task complexity (e.g., Zhang, 1987), and translation vs. direct writing (e.g., Kobayashi & Rinnert, 1992). These 16 studies used holistic scales (e.g., Hedgcock & Lefkowitz, 1992), error-free units (e.g., Casanave, 1994), error counts (Carlisle, 1989), and error-free counts with error classifications (e.g., Frantzen, 1995). She concluded that most of the studies provided too little information on the measures for anyone to replicate the studies. Furthermore, only half of the studies reported interrater reliability. This uncertainty of the reliability of the measures was problematic, not only because it made studies difficult to replicate but also because it called into question nonsignificant findings. Nonsignificant findings in a study can result from a variety of factors, one of which is unreliable measures. Given that researchers often report the results of previous studies, with no qualification, as evidence that a certain treatment had no effect, the lack of rigor in some studies is problematic.

Polio (1997) then conducted her own study on a set of 38 essays written in a one-hour timed writing by university students in an academic ESL program at a U.S. university. She used a holistic scale, three error-free measures (error-free T-units/total T-units, error-free T-units/total clauses, error-free T-units/total words), and an error classification system with 38 error-types also resulting in the number of errors per words. She found a range of reliability figures on the measures including very low reliability using the holistic scale and very high reliability using errors per words.

Recently, Evans, Hartshorn, McCollum, and Wolfersberger (2010) used FACET analysis to check the reliability of three measures (error-free T-units, error-free clauses, and weighted clause ratio, which takes into account error severity). In a FACET analysis, if the rater facet is close to 0, it indicates high rater agreement. They found the rater separation on the three measures to be 0.0–.02, which indicates excellent reliability.

In a review of studies on written corrective feedback completed after 1997, Polio (2012a) found that eight out of ten studies reported some type of reliability measure, which was an improvement from the studies described in her previous review. However, some of the studies calculated only intra-rater reliability, and some reported having more than one rater code only a portion of the data. Furthermore, none of the studies of written error correction included enough information for a replication, often because of lack of detailed information about the measures.

Validity

Polio (1997) mostly set aside the issue of validity, mentioning only that it needed to be investigated further and that all the measures might not be measuring the same construct. She did, however, include an extensive discussion of why raters might disagree on what an error is. For example, the issue of what is considered an error in academic writing vs.