Available online at www.sciencedirect.com

ScienceDirect

Journal of Second Language Writing 24 (2014) 33–50

JOURNAL OF SECOND LANGUAGE WRITING

Measuring written linguistic accuracy with weighted clause ratios: A question of validity

Norman W. Evans [a,*], K. James Hartshorn [b], Troy L. Cox [c], Teresa Martin de Jel [b]

[a] Brigham Young University, Department of Linguistics and English Language, 4050 JFSB, Provo, UT 84602, USA
[b] Brigham Young University, English Language Center, UPC, Provo, UT 84602, USA
[c] Brigham Young University, Center for Language Studies, JFSB, Provo, UT, 84602, USA

Abstract

Determining linguistic improvement in L2 writing requires a precise measure of linguistic accuracy. Although numerous metrics of linguistic accuracy have been used in L2 research, Wigglesworth and Foster (2008) proposed a new kind of measure—a weighted clause ratio—based on the adequacy of the writer's conveyed meaning. This paper evaluates the validity of this metric and compares it to two of the most similar measures of linguistic accuracy currently in use: the error-free T-unit ratio and the error-free clause ratio. The data collected and analyzed in this study were drawn from over 350 writing samples generated by 81 ESL writers whose language abilities range from low or intermediate to advanced. To provide baseline, comparative data, this study also analyzed writing samples from 16 native English-speaking students. This study utilized Many-Facet Rasch Measurement and other analyses to identify variables affecting the validity of the weighted clause ratio.

Keywords: Weighted clause ratio; Written corrective feedback; Linguistic accuracy; Communicative adequacy

Introduction

Teacher-administered error correction in students' second language writing has been a topic of considerable interest and some controversy over the past several decades, due principally to Truscott's (1996) provocative denunciation of the practice as futile and potentially ''harmful'' (p. 327). Since Truscott made that claim, many have countered that error correction is not only needed and expected by the learners but also, in some cases, a pedagogically sound practice (Bitchener & Knoch, 2009a, 2009b, 2010; Bitchener, Young, & Cameron, 2005; Bruton, 2009, 2010; Evans, Hartshorn, McCollum, & Wolfersberger, 2010; Evans, Hartshorn, & Strong-Krause, 2011; Ferris, 2002, 2003; Hartshorn & Evans, 2012; Hartshorn et al., 2010).

The central issue in the arguments for and against error correction in L2 writing has focused on whether or not writing improves as a consequence of corrective feedback. To some extent, the controversy surrounding this issue can be attributed to the difficulty of defining and measuring what is meant by *improvement* (Casanave, 2004). This

* Corresponding author at: Department of Linguistics and English Language, 4050 JFSB, Provo, UT 84602, USA. Tel.: +1 801 422 8472; fax: +1 801 422 0906.
E-mail addresses: norman_evans@byu.edu (N.W. Evans), james_hartshorn@byu.edu (K.J. Hartshorn), troy_cox@byu.edu (T.L. Cox), teresalovestravel@gmail.com (T. Martin de Jel).

http://dx.doi.org/10.1016/j.jslw.2014.02.005
1060-3743/

difficulty can be explained, in part, by the fact that writing is a multifaceted, complex process and product. Given this complexity, defining what constitutes improvement needs to be analyzed from various perspectives. For instance, though L2 writing used in authentic contexts is usually evaluated holistically for its overall communicative effect, it often is not evaluated by its constituent parts; however, important reasons do exist for analyzing L2 writing in terms of its discrete components such as fluency, complexity, rhetorical appropriateness, communicative adequacy, and linguistic accuracy (e.g., Evans et al., 2010; Pallotti, 2009; Skehan, 1998).

Since fluctuations in one component, such as fluency or complexity, might be associated with various fluctuations in another, such as accuracy (e.g., Bygate, 1999; Skehan, 1998, 2009; Skehan & Foster, 1997), each component of writing requires a separate and valid measurement to identify change. Without such independent measures, researchers will be limited in their ability to identify, track, and understand language development in L2 writing. Aiming to confirm the best general metric of linguistic accuracy, we attempt to validate one novel measure of linguistic accuracy—the weighted clause ratio (WCR) posited by Wigglesworth and Foster (2008).

While research must utilize specific accuracy measures for individual linguistic features if we are to fully understand language development (e.g., Norris & Ortega, 2009; Robinson & Ellis, 2008), research must also utilize an overall accuracy measure. Such a measure expands the researcher's understanding of the production as a whole, including the interconnected causes for errors that could be missed when examining only the accuracy of one or two linguistic features. Since this study's scope is limited to analyzing an overall measure of linguistic accuracy, we will not address measures designed to examine a single linguistic feature at one time, such as suppliance in obligatory context analysis (Brown, 1973), target-like use (Pica, 1983), or obligatory occasion analysis (Ellis & Barkhuizen, 2005).

Before proceeding, we also need to clarify two essential terms—*accuracy* and *error*. We define *accuracy* as "The ability to be free from errors while using language to communicate" (Wolfe-Quintero, Inagaki, & Kim, 1998, p. 33). Additionally, we define *error* 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).

Linguistic accuracy research: An overview

The quest for an appropriate way to measure linguistic accuracy in L2 writing is not new. Many measures have been devised over the years with various advantages and disadvantages. One relatively quick approach to measuring aspects of L2 writing has been the use of holistic or analytic scales. While such scales are the most efficient and least expensive to administer (e.g., Knoch, 2009; Shaw & Weir, 2007), most scales have not been designed to isolate linguistic accuracy and have often conflated various components of writing (e.g., Freedman, 1979; Hedgcock & Lefkowitz, 1992; McCarthey, Guo, & Cummins, 2005; Ojima, 2006; Tarone et al., 1993; Wesche, 1987).

While some studies have used a scale-based approach specifically to measure linguistic accuracy, many studies have not provided reliability statistics, making it difficult to interpret their findings (e.g., Barkaoui, 2010; Evans & Fisher, 2005; Lee, 2006; Lo & Hyland, 2007; Macaro & Masterman, 2006; Ruegg, Fritz, & Holland, 2011; Storch, 2009). However, a few studies of which we are aware have provided reliability information. For example, Hamp-Lyons and Henning (1991) used a 10-point scale (0–9) to analyze linguistic accuracy across writing tasks. Despite reporting reliability coefficients ranging from .70 to .79 for their Test of Written English, they achieved a range of only .33 to .35 for the Michigan Writing Assessment. This raises questions regarding the suitability of specific instruments across contexts. Polio (1997) also attempted to measure linguistic accuracy using a 12-point scale. Although she reported an intra-rater reliability of .77, inter-rater reliability ranged from .44 to .53, underscoring the reliability flaws associated with rater training when using scale-based approaches.

Two additional studies reported higher inter-rater reliability statistics for linguistic accuracy measures. The first was Sasaki (2000), who used the *language use* section of the ESL Composition Profile (Jacobs, Zinkgraf, Wormouth, Hartfiel, & Hughey, 1981). She reported an inter-rater reliability correlation of .88. The second study was Stevenson, Schoonen, and de Glopper (2006), who reported a .90 correlation, but they provide no information regarding the scale or its criteria. Scholars such as Weigle (2002) have pointed out that, with enough training, researchers can observe high levels of reliability; however, it is unclear if this is due to raters' improved use of the instrument or simply an increased familiarity with each other from working together on multiple projects. Consequently, if rubrics cannot be used reliably across different raters and contexts, they may not be well suited for careful research. Such concerns have led