CrossMark

Annual Review of Applied Linguistics, **36** (2016), pp. 98–116.
© Cambridge University Press, 2016
doi: 10.1017/S0267190515000082

Capturing Accuracy in Second Language Performance: The Case for a Weighted Clause Ratio

PAULINE FOSTER
St. Mary's University
pauline.foster@stmarys.ac.uk

GILLIAN WIGGLESWORTH
University of Melbourne
g.wigglesworth@unimelb.edu.au

ABSTRACT

As increasing numbers of research papers in applied linguistics, language learning, and assessment use discourse analysis techniques to assess accuracy in performance, it is timely to examine at a detailed level the wide variety of measures employed. Ideally, measures need to capture accuracy in as valid and reliable a way as possible, but this has proved elusive. In this article, we systematically review the variety of different measures in used in these fields, both global and local, before presenting a more finely tuned weighted clause ratio measure which classifies errors at different levels, that is, those that seriously impede communication, those that impair communication to some degree, and those that do not impair communication at all. The problem of reliably identifying these levels is discussed, followed by an analysis of samples from written and spoken second language performance data. This new measure, grounded in a comprehensive review of prior practice in the field, has the advantages of being relatively easy to use, measuring accuracy rather than error, and evaluating smaller increases in improved performance than have previously been possible.

INTRODUCTION

The point at which a learner can be said to have acquired a given feature of a second language (L2) is not something applied linguists agree upon. Depending on the preferred theoretical perspective, acquisition can be variously described, as Norris and Ortega (2003) pointed out, as the ability to do the following:

- State the particular rule or define the particular word.
- Use that rule or that word fluently in a variety of contexts and on a number of occasions.
- Recognize non-native-like examples of the word or the rule.

What the various perspectives might agree on, however, is that in the long view, the route to acquisition is characterized by movement towards accuracy in performance. This might ascribe increasing accuracy to the learner's access to universal grammar, the length and breadth of exposure to the L2, positive engagement with

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

the L2 community, linguistic aptitude, moments of psychological insight into the structure of the L2, or some combination of these factors. From whichever standpoint, increasing accuracy in L2 performance suggests developmental progress.

For both L2 learning researchers, and applied linguists more generally, how accuracy in performance develops over time and how it changes between contexts or according to a variety of different variables (e.g., planning time, task familiarity, and different tasks) mean that a reliable and nuanced measure of accuracy is crucial for the field and for moving forward our understanding of improving accuracy. Over the years, a variety of instruments to measure accuracy have been employed, from error analysis and obligatory occasion analysis to analysis of target-like usage of specific morphemes to, more recently, the incidence of error-free clauses. The early tendency to omit reference to inter- and/or intrarater reliability (see Polio, 1997) has now largely been rectified, but the measures themselves remain very disparate. In our view, all are problematic to some degree. In the next paragraphs, we discuss the variety of options that have, to date, been put forward.

By way of illustration of the numerous accuracy measures in use, we take a 2007 volume, edited by Van Daele, Housen, Kuiken, Pierrard, and Vedder, of collected studies on complexity, fluency, and accuracy in L2 performance and describe the measures chosen by the contributors who present accounts of accuracy. We begin with Granfeldt's (2007) study, which looked at accuracy in L2 French. Granfeldt calculated errors per T-unit for both oral and written data, refining these further into two types: lexical errors per T-unit and grammatical errors per T-unit. Gunnarsson (2007) also looked at accuracy in the written production of L2 French, but chose the specific measures of subject-verb agreement in present tense of *-ir*, *-re*, and *-oir* verbs, and correct use of the *passé compose* and *imparfait*. Lee and Oh (2007), investigating the effects of pretask planning on L2 English oral narrative tasks, applied the measure of error-free T-units and number of errors per T-unit to their data. However, their Korean first language (L1) participants were of low proficiency and produced almost no error-free T-units. To cope with this outcome, Lee and Oh decided to count as error-free those units with no errors *and* those with one error. Furthermore, they did not code missing or unnecessary definite or indefinite article as errors at all, because these were so pervasive and "could override the difference in other types of error" (p. 130). Lee, Joo, Moon and Hong (2007) also using oral data, investigated the effects of interlocutor type on English L2 performance. Accuracy was measured in three ways: through lexical errors per 100 words, morphosyntactic errors per 100 words and also error-free T-units, again counting a T-unit with one error as error-free. Roquet-Pugès and Pérez-Vidal, (2007) looked at the acquisition of competence in writing in a third language, measuring accuracy through errors per clause and percentage of error-free clauses. Skehan and Foster (2007), reporting a meta-analysis of studies into task-based English L2 oral data, reported accuracy through percentage of error-free clauses, supplemented by another measure that (unusually) takes into account the disproportionate effect on accuracy scores of short clauses. Given that short clauses are more likely to be error-free than longer ones, Skehan and Foster proposed measuring accuracy by looking at the length of error-free run, that is, the number

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

of words for which clauses are at least 70% accurate. Tonkyn (2007) investigating short-term changes in accuracy in oral data, employed a whole battery of measures: number of words divided by number of errors, number of error-free AS-units divided by total units, number of words divided by number of verb phrase errors, and number of words divided by number of noun phrase errors. Van Daele, Housen, Pierrard, (2007), investigating accuracy in the development of two L2s, calculated a ratio of lexical accuracy and grammatical accuracy by dividing the total number of both kinds of error by total clauses. Finally, Kuiken and Vedder (2007) reported a study of written L2 in which they measured accuracy as total number of errors per T-unit. Uniquely in this volume, Kuiken and Vedder discriminated between first-degree errors (minor deviations in spelling, meaning, or grammatical form that do not interfere with the comprehensibility of the text), second-degree errors (which are more serious), and third-degree errors (which make the text "nearly incomprehensible") (p. 117).

Thus, inside this one collected volume—Van Daele, Housen, Kuiken, et al. (2007)—we have a useful snapshot of the very wide diversity in the measurement of accuracy in L2 data and the problems that this diversity continues to engender. The somewhat porous boundaries of the definition of the term error-free are obvious. We see that Lee and Oh (2007) decided to make the term error-free fit their particular data by deeming inaccurate use of articles to be the same as accurate use of articles, while Lee et al. (2007) tweaked things by deciding that a clause with just one error, of whatever kind, is equal to a clause with no error at all. We hasten to say that this is not meant as a criticism of the researchers; they have, after all, recognized the shortcomings of the measure and, *faut de mieux,* tried to make it workable for their data. But designing nonce solutions for particular data sets is not a satisfactory state of affairs for the field as a whole. However carefully worked out, nonce solutions inevitably compromise the generalizability of research findings and diminish the scope for meta-analyses. In a less ad hoc fashion, we see that Kuiken and Vedder (2007) proposed a way round the problem of the insufficient discrimination of the error-free clause instrument by categorizing errors as minor, middling, or major; in effect, they were opting for the simultaneous use of three scores for accuracy. Similarly, Skehan and Foster (2007) acknowledged the related problem of very short clauses emerging as disproportionately accurate when compared with longer ones, and so proposed a measure that takes length of clause into account. Nevertheless, as Ellis and Barkhuizen (2005, p. 366) noted, the proliferation of measures trying to capture the same aspect of L2 performance means that robust comparisons between studies is impossible, and this undermines any attempt to construct a bigger picture from their results. This is a fundamental difficulty for the field. (See also Plonsky & Kim, this volume.)

MEASURES OF L2 ACCURACY

In this section we discuss in some detail the definitions and applications of the most commonly used measures of accuracy in L2 performance. They divide into two

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

broad categories through a simple distinction: Local measures aim to track the use of designated grammatical features in the L2 such as verb and noun morphology, while global measures focus on gauging the overall level of accuracy in an oral or written performance. Taking local measures first, the most commonly used are *supplied in obligatory context* (SOC), *target-like use* (TLU), and *used in obligatory contexts* (UOC).

SOC is applied by counting all the occasions in an L2 text or transcript where the chosen grammatical feature is both obligatory and correctly supplied, and then dividing the total by the number of contexts where the feature is obligatory. Perfect accuracy is thus a score of 1.00, representing correct use on every necessary occasion. The drawback here is that learners are apt to use a grammatical feature when it is not obligatory (as in **they goes to the cinema*) and SOC cannot pick up oversuppliance. An SOC score of 1.00 therefore does not necessarily reflect perfectly accurate use of the feature under investigation, and in fact it is likely to be an overestimate.

TLU does take account of oversuppliance. This measure is calculated by counting the number of correct suppliances of the feature in obligatory contexts, and dividing this by number of obligatory contexts, *plus* the number of suppliances in nonobligatory contexts. Thus, the more often the learner supplies a grammatical feature where is it not needed, the greater the denominator and the lower the TLU score, even if every use in obligatory contexts is in fact perfectly right. If the learner gets the feature right in obligatory contexts *and* never supplies it where it is not needed, then the score will be a perfect 1.00.

UOC is the accuracy measure that gauges the degree of overuse of a grammatical feature. It is calculated by adding together the number of suppliances in both obligatory and nonobligatory contexts, then dividing this by the number of obligatory contexts. A score over 1.00 reflects the degree to which the learner overuses the feature.

Local measures such as these may be appropriate for measuring the development of specific morphological features across time, especially if the research study is focused upon, say, the treatment of this feature through different pedagogical interventions. However, as Norris and Ortega (2003), among others, argued, development in control of an L2 structure is not necessarily linear, so conclusions about its progress at any given point in time are not straightforward. (In fact, Wolfe-Quintero, Inagaki, & Kim, 1998, described a more omega-shaped curve in the acquisition of L2 morphemes.) In other words, if a learner can be shown to improve in the use of the targeted feature, it can be extrapolated to an interlanguage improvement across the board. But Ellis and Barkhuizen (2005, p. 59) argued that the validity of local measures remains problematic because "learners do not acquire grammatical features concurrently. Rather, some features are acquired early and others late."

Hence, a researcher using a local measure to reflect a learner's overall L2 development faces the knotty problem of having to decide which grammatical feature will provide its best reflection. Also, (from a purely practical perspective) the researcher must be confident that the chosen feature will occur often enough in the

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

data to make its measurement meaningful. It is useless to track the correct use of irregular plural nouns if these turn up infrequently.

Equally, it is necessary to decide what consideration should be given to the influence of possible L1 transfer. Tracking the development of the English article system in learners from an L1 background where the L1 does not have articles runs the risk of relying on something which is nonsalient, frequently redundant, and susceptible to early fossilization. On balance, therefore, we would argue that local measures of accuracy are most valid in circumstances where the focus is on how development of a grammatical morpheme responds to particular treatments.

Global measures, by contrast, examine the text or transcript in its entirety. They essentially work by dividing the data into segments and calculating the error rate. The segmentation may be made by dividing the data into units of, say, 100 words each (*mean errors per hundred words*) or into syntactic units of some kind (*mean errors per AS-unit, per T-unit, per clause*). Another way to calculate accuracy this way is to measure the proportion of the units that are error-free (*percentage of error-free clauses, T-units, or AS-units*). But whereas the foundation of local measures rests on a somewhat easier question—*What do we mean by accurate and nonaccurate use?*—the foundation of global measures rests on two trickier questions: *What do we mean by a word, a clause, T-unit, and how do we count errors?*

It is not always a simple matter to draw word boundaries. Are *wineglass* or *teacup* single words or not? What about *wine bar* and *tea time*? Orthographic conventions are of no use here as our focus is psycholinguistic reality. Contractions such as *isn't, can't,* and *mustn't* are surely selected from memory as wholes, as are the vast array of conventional chunks of language such as *how are you, as a matter of fact, fast and furious,* and *on top of the world.* Are these also to be counted as one word*?* There is no easy answer to such a question, but even if we could arrive at a satisfactory definition of what counts as one word, a segmentation unit of 100 words still needs justification. Why choose 100 words? Which 100 words? From which section of the transcript? Would 120 or 80 be better?

Segmentation of units at syntactic boundaries does have a claim to psycholinguistic reality (Foster, Tonkyn, & Wigglesworth, 2000), rendering units such as a clause, a T-unit, and an AS-unit stronger candidates for the basis of an analysis. The T-unit (Hunt, 1965) was designed for the analysis of written language and approximates to a sentence. Though these have been, and continue to be, used in the analysis of the spoken language (as in Granfeldt, 2007, discussed further above), the pervasively fragmentary nature of speech makes this tricky. The communication unit, or c-unit (Loban, 1966), was designed specifically to cope with fragmentary oral data and is widely used (e.g., Foster & Skehan, 1996; Lin & Hedgcock, 1996) in spite of being rather vaguely defined. Building on this, Foster et al. (2000) provided a definition of an analysis of speech unit (AS-unit) with the level of detail required by anyone attempting to segment spoken language. They further defined in detail the *clause* as a syntactic subunit, allowing a more finely grained analysis of oral data by basing it on something that is likely to be shorter, and thus less prone to error, than a complete unit.

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

Once the basic unit of measurement has been chosen, and the data have been reliably segmented, there is a choice between calculating the *mean number* of errors per unit or simply identifying the *whole unit* as either error-free or containing error(s). While this procedure is arithmetically simple, the identification of errors in L2 data is far from that. Essentially, error identification is the process of editing an L2 performance into something that has no errors. But as any teacher or researcher knows, there is more than one way to correct errors, and some involve fewer changes than others. Example 1 is from an L2 language learner's written text and comprises one T-unit:

1. *Nowadays, we have to accept that education is the best way of people who want to be the good situation in future.*

The corrected Examples 2–5 below serve to illustrate that rendering this error-free can be done in many ways. (Deleted words are crossed through and supplied corrections are underlined.)

2. *Nowadays, we have to accept that education is the best ~~way of~~ choice for people who want to be ~~the~~ in a good situation in the future.*
3. *Nowadays, we have to accept that education is the best way ~~of~~ for people who want ~~to be in~~ a good situation in the future to achieve their goals.*
4. *Nowadays, we have to accept that education is the best way ~~of~~ for people ~~who want~~ to be ~~the~~ in a good situation in future.*
5. *Nowadays, we have to accept that education is the best ~~way~~ investment for people who want ~~to be in~~ a good ~~situation~~ life in the future.*

How many errors have been identified in Example 2? Is changing "way of" to "choice for" counted as one error, or two? Is changing "the" for "in a" one error or two? Have three errors been identified in Example 3? Or does deleting three words in a row count as three individual errors, making a total of five? Does Example 4 represent the correction of three, four, or five errors? Does Example 5 show four, five, six, or seven corrected errors?

This brief example of how a single sentence may be corrected in five different ways raises obvious concerns about the reliability and validity of an error count. A solution is to base an analysis on syntactic units such as clauses, T-units, and AS-units, which can be more reliably identified and designated as error-free or not error-free. In this way, global accuracy is expressed as the percentage of error-free units to total units. As we have noted above, however, the drawback is that the longer the unit, the greater the likelihood that the learner will make an error, thereby unfairly disadvantaging more complex units that entail multiple clauses. This is partially mitigated by choosing clauses as the basis for analysis rather than full AS-units or T-units, even though clauses themselves can also be protracted and more likely to include an error as a result. An additional problem is that a simple binary distinction in coding units as either error-free or not error-free means that a unit which is not error-free is scored the same no matter what error(s) it contains.

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

Examples 6 and 7 illustrate this (with the clause boundaries indicated by double colons).

A B

6. *In multiple-choiced exams, it is not hard :: to make an educated guess, or a random one.*

C *D*

7. *On the other side of the story, I see :: their mum see the childs a big map*

Example 6 is one T-unit of written L2 English, Example 7 is one AS-unit of spoken L2 English. In each case we have a unit that contains two clauses. In Example 6 the first clause (A) has a mistaken letter *d* affixed to *multiple-choice*, while the subsequent clause (B) has no errors. In Example 7 the first clause (C) has the word *story* wrongly selected for *picture*, while the subsequent clause (D) has multiple errors: vocabulary choice (*see* for *show*), verb morphology (*show* for *is showing*) and noun inflection (*childs* for *children*). In an error-free clause analysis, only (B) is error-free, and clause (D), which is almost unintelligible, is graded the same as clause (A), where the error is practically unnoticeable, and as clause (C), where the error is noticeable but the meaning is not compromised.

From this review, we conclude that for most research purposes, global measures based on a syntactic unit can provide a better evaluation of accuracy in L2 performance than can local measures, and the best tool in the measurement kit is currently the error-free clause because it combines a reliably defined and valid unit with a finer-grained analysis than offered by a whole T-unit or AS-unit analysis. However, the fact that in a single clause one small error carries the same weight as several major errors means that finely grained differences in accuracy will remain below its radar.

WEIGHTING ERRORS

Addressing the problem of an almost accurate clause being graded as equal to one filled with errors requires a measure that takes gravity of error into account. Such an idea is not new. As long ago as 1975, Nas (1975) defined first-, second-, and third-degree errors in L2 writing, relating them to levels of comprehensibility; first-degree errors offer no impediment to comprehensibility, second-degree errors impede comprehensibility to some extent, and third-degree errors compromise it seriously. Similarly, and as noted above, Kuiken and Vedder (2008, p. 53) sorted errors in L2 written data into three levels according to their impact on communicative success or adequacy. The measures of both Nas (1975) and Kuiken and Vedder (2008) produced information about the proportion of data characterized by each of the three levels of error severity, but—crucially—did not combine them into a global score for the data as a whole. Two further drawbacks are that both

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

measures were designed for written data using a whole T-unit analysis rather than a finer-grained clause analysis, and neither has been tried out on AS-units of spoken data. The challenge then is to design a new instrument that produces a weighted clause ratio (WCR) that can be applied to both written and spoken clauses and will deliver a global score.

Before going further into the design of a new measure, it is important to discuss the central issue of how errors can be reliably categorized as first, second, and third degree. Both Nas (1975) and Kuiken and Vedder (2008) invoked communicative adequacy in determining error gravity. Pallotti (2009) discussed this construct as distinct from accuracy, pointing out that grammatically accurate language can be entirely inadequate for successful communication, while grammatically inaccurate language can deliver a message perfectly well.

> If, in an information gap task, a learner were to utter unhesitatingly *colorless green ideas sleep furiously on the justification where phonemes like to plead vessels for diminishing temperature,* her production would score extremely high on [complexity, accuracy and fluency], in spite of being completely irrelevant and probably counter-productive for task success. In contrast, an utterance such as *No put green thing near bottle. Put under table* is neither complex not accurate, and may not be fluent either, but can turn out to be perfectly functional for achieving the speaker's (and the task's) intended communicative goal. (Pallotti, 2009, pp. 596–597)

In this view, judgments of communicative adequacy can have little to do with measurements of grammatical accuracy as the two constructs are not necessarily related. Indeed, Pallotti (2009) noted that it is possible to regard communicative adequacy as theoretically independent from accuracy. Yet from a more practical perspective, and because research participants have ordinary instincts for communication, an analysis of spoken or written data should be able to start from the assumption that the speaker or writer is *not* trying to be deliberately obscure, irrelevant, untruthful, or ambiguous, but is following Grice's (1975) maxims and cooperative principle to convey a clear message. If the old chestnut of *colorless green ideas sleep furiously* were ever to show up in a transcript, one would expect it to be in the context of a question like, *What was Chomsky's famous example of a sentence that is grammatical but makes no sense*? As such, it would be both grammatically accurate and communicatively adequate. At the other end of the scale, while it is true that highly *in*accurate language can deliver a message, it might not be communicatively adequate without the speaker relying on the immediate context to supply information about, for example, person, number, tense, mood, and aspect. Without such information, the intended message, *He couldn't have put the green vase next to those bottles because she had already hidden it under the desk,* would not be adequately communicated by *No put green thing near bottle. Put under table.*

Communicative adequacy can be achieved using gesture alone. A person with no knowledge of a language can, for example, get a waiter to refill her water glass by establishing eye contact, smiling, pointing at the empty glass, and lifting it in

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

TABLE 1. *Categorizing Clauses for Accuracy*

Clause Category	Definition
Entirely accurate	The clause is accurately constructed.
Level 1	The clause has only minor errors (e.g., in morphosyntax) that do not compromise meaning.
Level 2	The clause contains serious errors (e.g., verb tense, word choice, or word order), but the meaning is recoverable, though not always obvious.
Level 3	The clause has very serious errors that make the intended meaning far from obvious and only partly recoverable.

a drinking motion. But as language-less communication strategy works only by recruiting general knowledge and the immediate context, it can't take communication very far. To go further than this, language must be deployed, and the more layered the meaning to be conveyed, the more grammatical and lexical knowledge is required to transmit it. In this regard, a language learner with developing L2 knowledge who does not wish to be bound to the here-and-now in which pidgins operate is moving from a position of not being able to express messages efficiently and effectively toward a position of being able to do all this. But many language learners, of course, go further still and arrive at a position of using grammatical morphemes that add nothing at all to meaning, such as the English third-person, present-tense verb ending. This shows language proficiency beyond adequacy, so it is necessary for an accuracy measure to be able to take into account what might be fully adequate but still inaccurate.

Using a Weighted Clause Ratio Measure: Categorization and Rating

In this section, we turn to the detail of the practical application to spoken and written L2 data of a WCR. There are three steps: clause boundary identification, clause categorization, and clause rating. Our approach follows Nas (1975) and Kuiken and Vedder (2008) in assigning clauses to three levels depending on their error gravity, though our ultimate aim is different in that we use these error categories to extract and measure what is accurate.

Clause boundary identification is achieved by a T-unit or AS-unit analysis for written and spoken data, respectively, following Hunt (1965) for the T-unit and Foster et al. (2000) for the AS-unit. Both types of analyses are based on syntax, with unit boundaries falling where a syntactic structure is complete, and subunit boundaries (i.e., clauses) falling within them.[1] Once this is done, the next step is to categorize each clause according to the gravity of error within it. Table 1 above shows five categories a clause may be assigned to, ranging from the entirely accurate to the entirely inaccurate.

It goes without saying, we hope, that clauses with serious error are likely to include other errors which are less serious, so effectively a clause is categorized according to its worst-level error, whatever that might be. As Table 1 shows, the

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

TABLE 2. *Scoring the Categories*

Clause Category	Score
Entirely accurate	1.00
Level 1 error	0.80
Level 2 error	0.50
Level 3 error	0.10

most serious errors are those that compromise comprehensibility; an L2 performance characterized by these reflect beginner or low levels of L2 proficiency. Less serious errors are those that impede but do not compromise comprehensibility; a performance characterized by these reflects higher levels of L2 proficiency. The least serious errors are those that do not impede comprehensibility at all; a performance which is characterized by these reflects even higher levels of L2 proficiency. A performance characterized by only occasional Level 1 errors and none at all of Level 2 or 3 would reflect a very high L2 proficiency. There is a developmental arc; lower-level learners are prone to more serious, meaning-compromising errors compared to higher-level learners, and highest-level learners avoid even the errors that do not impede meaning at all.

Leaving aside the question of eloquence, we argue that it is generally straightforward to identify a clause that has no error in morphosyntax or word choice, and it is generally straightforward to identify a clause with so many such errors that it is entirely incomprehensible.[2] Between these are three categories of clause with different levels of error that need more thought in assignment. Clauses at Level 1 contain only minor errors (such as omission of "s" in the third-person singular in the English present tense) that do not impact on comprehensibility. Clauses at Level 3 have errors that do impact comprehensibility, rendering the intended meaning difficult to recover. Clauses at Level 2 sit between these; their meaning is impacted by error but not derailed by it. (To illustrate further, we are providing in Appendices A and B a sample of oral and written data divided into clauses and categorized according to level of error.)

Each level is scored to reflect the accuracy of the clause. Taking the entirely accurate clause as a starting point, we give this an accuracy score of 1.00. For clauses categorized as Level 1, we give an accuracy score of 0.80. For clauses categorized as Level 3, we give an accuracy score of 0.10, crediting the fact that though it has serious errors, it nevertheless has some accurate language within it. Finally, for clauses categorized as Level 2, we give an accuracy score of 0.50. The system is summarized in Table 2.

We arrived at this particular calibration after careful trialing on a variety of spoken and written data, and found this to be the one that best distinguished L2 outcomes along a range of implementation variables such as task repetition and pretask planning, and across time. The final step in measurement is to produce an index of accuracy by adding up the total clause ratings for a transcript and dividing the sum by the total number of clauses. This produces a mean accuracy rating for

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

TABLE 3a. *Profile 1: Advanced Writing*

	Time 1	Time 2
T-units	20	23
Clauses	39	42
Complexity (total clause/total T-units)	1.95	1.83
Entirely accurate T-units	13	17
Entirely accurate clauses	31	37
Clauses at Level 1	8	5
Clauses at Level 2	0	0
Clauses at Level 3	0	0
Weighting		
Entirely accurate clauses × 1.00	31.00	37.00
Level 1 × 0.80	6.40	4.00
Level 2 × 0.50	0.00	0.00
Level 3 × 0.10	0.00	0.00
Raw total	37.40	41.00
Weighted Clause Ratio (raw total divided by total clauses)	**0.96**	**0.98**

the transcript as a whole, where 1.00 would represent a completely accurate L2 performance.

Practical Illustrations

In the following section, we illustrate this new unit of measurement in more practical detail, presenting two writing profiles and two speaking profiles using authentic data collected at different points in time.

Written Data. The written data were taken from a study of writing feedback (see Storch & Wigglesworth, 2007, 2010a, 2010b; Wigglesworth & Storch, 2009) in which data were collected from students writing individually and in pairs. For this analysis, only the individual data were used. In the study, at Time 1 the students wrote an argumentative essay on the role of education in future success. Five days later, they were provided with feedback on the accuracy of the essay in the form of editing symbols. They were given 15 minutes to study the feedback, which was then removed, and they then rewrote the essay. In the analysis below, the performances at Time 1 and Time 2 of one advanced-level and of one intermediate-level student are analyzed using the variety of different measures discussed above, as well as the new measure we present here.

The first profile comes from an advanced student who was enrolled in a university-level undergraduate course, and who had achieved an IELTS[3] of at least 6.5. Table 3a provides the profile of the student's writing when comparing the essay at Time 1 with the essay at Time 2, five days later, following feedback.

In this profile, we see a very good performance at Time 1 improving slightly at Time 2. The second performance is longer by three more clauses (42 compared to 39) and has three fewer clauses at Level 1. There are no clauses at Level 2 or 3 in either performance. The complexity index (achieved by dividing the number of

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

TABLE 3b. *Menu of Accuracy Scores for Profile 1*

	Time 1	Time 2
% Error-free T-units	65.00	73.91
% Error-free clauses	79.48	88.09
Errors per 100 words	3.20	2.11
% TLU of plurals	98.00	94.74
% TLU of articles	100	100
Weighted Clause Ratio (raw total divided by total clauses)	**0.96**	**0.98**

TABLE 4a. *Profile 2: Intermediate Writing*

	Time 1	Time 2
T-units	17	20
Clauses	35	41
Complexity (total clause/total T-units)	2.06	2.05
Entirely accurate T-units	2	2
Entirely accurate clauses	4	11
Clauses at Level 1	13	13
Clauses at Level 2	13	12
Clauses at Level 3	5	5
Weighting		
Entirely accurate clauses × 1.00	4.00	11.00
Level 1 × 0.80	10.4	10.4
Level 2 × 0.50	6.5	6.0
Level 3 × 0.10	0.5	0.5
Raw total	21.40	27.90
Weighted Clause Ratio (raw total divided by total clauses)	**0.61**	**0.68**

clauses by the number of T-units) is also given, and shows that the performance at Time 2 was marginally less complex. The WCR scores of 0.96 for Time 1 and 0.98 for Time 2 capture this profile of slight improvement.

When we compare in Table 3b the new measure to the two other kinds of global measures discussed above, we find that both error-free T-units and error-free clauses report a much larger improvement in accuracy at Time 2, while errors per 100 words show a decrease. For the local measures, target-like use of plurals has also decreased slightly at Time 2, and TLU of articles remains the same at 100%.

The second profile we present here in table 4a is from an intermediate-level student who was studying English for the purposes of entering into a university course who had begun with an IELTS between 5.5 and 6.5.

This student is less proficient in English L2 compared with the one shown in Tables 3a and 3b, producing a considerable number of clauses with Level 2 and 3 errors on both performances. Her improvement at Time 2 is associated with a larger number of clauses overall (41 compared to 35), and an increased number of clauses without error (11 compared to 4), with her complexity remaining the

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

TABLE 4b. *Menu of Accuracy Scores for Profile 2*

	Time 1	Time 2
% Error-free T-units	11.76	10.00
% Error-free clauses	11.43	26.83
Errors per 100 words	18.62	15.47
% TLU of plurals	72.00	63.33
% TLU of articles	33.33	34.78
Weighted Clause Ratio (raw total divided by total clauses)	**0.61**	**0.68**

same. The incidence of clauses with more grave errors does not change. So Profile 2 shows a small improvement at Time 2. Looking at the other measures for this performance, shown in Table 4b, we find that the rate of error-free clauses suggests quite a reasonable improvement at Time 2 (26.83% compared to 11.43%) while the rate of errors per 100 words suggests a much smaller one. The TLU of plurals shows a worse performance, and the TLU of articles shows no change.

In summary, for Profile 1 the WCR measure reflects a shift from a very good piece of L2 writing to a very slightly better one at Time 2. For Profile 2, the WCR measure shows a small improvement due to a few more error-free clauses in an otherwise very similar performance. The other accuracy measures calculated here are contradictory in their indications.

The Oral Data. In contrast to the written data presented here, the oral data were not collected as part of a wider study, but specifically to trial the application of a WCR. Six upper-intermediate to advanced learners of English as a foreign language were individually recorded recounting a narrative told in a wordless cartoon strip (Sempé, 1973). After the first performance, they were asked to repeat the task, and then to repeat it a third time. The three performances represent the same person doing the same task within the space of 5 to 6 minutes. They were given general encouragement between the performances, but no specific feedback. Table 5 presents the profile of one upper-intermediate student across the three narrations of the story.

Across these task repetitions, we see no change in the relatively small number of the Level 2 and 3 clauses. There is a gradual increase in the number of entirely correct clauses and a gradual decrease in the number of those at Level 1. The fact that the data were collected from the same person in the space of just a few minutes would not lead us to expect a large improvement, and this is what the WCR reflects: a fairly accurate performance at Time 1 improves by Time 3 to a more accurate one as Level 1 clauses give way to those with no error. The error-free clause score, by comparison, suggests accuracy across performances has more than doubled, and the TLU of articles shows also a large improvement from Time 1 to Time 3, which could be the source of most of the increased accuracy here. The incidence of clauses at Levels 2 and 3 remains the same, suggesting that the more serious errors were not affected by repetition.

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

TABLE 5. *Profile 3: Upper-Intermediate Speaking*

	Time 1	Time 2	Time 3
AS-units	15	13	14
Clauses	22	22	19
Complexity	1.46	1.70	1.36
Entirely accurate clauses	5	9	11
Clauses at Level 1	12	9	6
Clause at Level 2	3	2	1
Clauses at Level 3	2	2	1
Weighting			
Entirely accurate × 1.0	5.0	9.0	11.0
Level 1 × 0.8	9.6	7.2	4.8
Level 2 × 0.5	1.5	1.0	1.0
Level 3 × 0.1	0.2	0.2	0.1
Raw total score	16.3	17.4	16.9
Weighted Clause Ratio (raw total divided by total clauses)	.74	.79	.89
% Error-free clauses	22.72%	41.09%	57.89%
%TLU of articles	60.00%	70.37%	83.33%

TABLE 6. *Profile 4: Advanced Speaking*

	Time 1	Time 2	Time 3
AS-units	14	13	18
Clauses	16	19	22
Complexity	1.14	1.46	1.22
Entirely accurate clauses	7	16	19
Clauses at Level 1	3	2	1
Clause at Level 2	4	1	2
Clauses at Level 3	0	0	0
Weighting			
Entirely accurate × 1.0	7	16	19
Level 1 × 0.8	2.4	1.6	0.8
Level 2 × 0.5	2.0	0.5	1.0
Level 3 × 0.1	0.0	0.0	0.0
Raw total score	11.4	18.1	20.8
Weighted Clause Ratio (raw total divided by total clauses)	.71	.95	.95
% error-free clauses	43.75%	84.21%	86.36%
%TLU of articles	83.3%	92.34%	91.75%

Profile 4 is of a speaker whose English was very fluent and assured. Her first performance nevertheless shows a number of clauses at Levels 1 and 2, which she reduces at Times 2 and 3. She produces more language at Times 2 and 3 and increases by quite a margin the proportion of this that is error-free. The error-free clause score mirrors the WCR scores in showing a great improvement at Time 2,

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

sustained at Time 3, though the WCR gives perhaps a better picture of Time 1 for which error-free clauses indicates a rather low score.

DISCUSSION

In contrast to the simple binary distinction of accurate and inaccurate, this WCR allows researchers considerably more insight into the language performance of students by identifying different levels of error, and thus allowing L2 data that are not entirely accurate to be nevertheless accounted for along a cline of error gravity. By sorting oral or written data into error categories, the WCR potentially offers researchers a window into the kinds of errors learners make at different levels of proficiency, at different times, and in different conditions. It appears sensitive to even small changes in accuracy, which other measures do not reflect. Uniquely, its weighted scoring mechanism produces a single index of accuracy, allowing easier cross-study comparisons than are possible now. By giving credits for accuracy rather than debits for error, it provides researchers with a more complete account of the movement of the learner towards more accurate performance in an L2.

In its draft form, the WCR was trialed by Evans, Hartshorn, Cox, and de Jel (2014) on written data and found to be robust compared to other measures, though the authors reported it as relatively time-consuming to implement. In our experience, however, as with most analysis, accumulated practice and familiarity speeds things up considerably. Once clause boundaries have been identified, the application of the measure becomes a relatively simple matter and allows significant reliability indices to be reported quickly. In a recent study that uses the measure to compare accuracy of L2 performance across two conditions, (Foster, 2016) a large sample of oral data was coded for accuracy independently by two researchers and returned an interrater reliability score of over 90% on all clause levels. The boundary between a Level 1 and a Level 2 error was the most prone to disagreement but these were invariably resolved through discussion.

We recognize that the impact of error on meaning is often felt beyond clause boundaries. Anyone who has worked on assessing accuracy in L2 data will know this only too well; some degree of personal judgment has to be invoked occasionally. And we must caution that, in keeping with other measures of accuracy, the WCR is interpreted best in the company of a measure of complexity. It is easier to be accurate in an L2 when being conservative in what you attempt to express. In spoken language in particular, which is under the pressure of real-time processing, the syntactic breathing spaces provided by clause boundaries invite a performance characterized by chaining simple clauses. In both the oral and written data used here as illustrations, we have provided an index of complexity to aid interpretation. Our illustrations of L2 performances in the tables above and also in the appendices demonstrate the important point that, in common with all measures of accuracy discussed in this article, a WCR ratio is designed to track morphosyntactic and

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

semantic accuracy rather than idiomaticity or style. Accurate writing and speaking can be wooden and clumsy, but these are problems that are not confined to L2 learners.

To conclude, we believe that a WCR, as we define and implement it here, avoids many of the problems of other measures on offer (although certainly not all of them). It is relatively simple to use and sensitive to very small changes in accuracy, and we hope these characteristics will recommend it to researchers in task-based L2 language performance.

NOTES

1 Foster et al. (2000) have copious examples of unit and clause boundaries.

2 Total incomprehensibility is quite rare in our experience. There is nearly always something in an otherwise incomprehensible clause which bears some meaning.

3 IELTS is the British International English Language Testing System used to assess language proficiency (similar to TOEFL).

REFERENCES

Ellis, R., & Barkhuizen, G. (2005). *Analyzing learner language*. Oxford, UK: Oxford University Press.

Evans, N. K., Hartshorn, K. J., Cox, T. L., & de Jel, T. M. (2014). Measuring written linguistic accuracy with weighted clause ratios: A question of validity. *Journal of Second Language Writing*, *24*, 33–50.

Foster, P. (2016). *Comparing classroom and laboratory task performance: An empirical study.* Manuscript in preparation.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*, 299–323.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). A unit for all measures: Analyzing spoken discourse. *Applied Linguistics, 21*(3), 354–375.

Granfeldt, J. (2007). Speaking and writing in L2 French: Exploring effects on fluency, complexity and accuracy. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning & teaching* (pp. 87–89). Brussels, Belgium: KVAB.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Studies in syntax and semantics III: Speech acts* (pp. 183–198). New York, NY: Academic Press.

Gunnarsson, C. (2007). Fluency and accuracy in the written production of L2 French. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning & teaching* (pp. 99–112). Brussels, Belgium: KVAB.

Hunt, K. (1965). *Grammatical structures written at three grade levels*. Champaign, IL: National Council of Teachers of English.

Kuiken, F., & Vedder, I. (2007). Task complexity, task characteristics and measures of linguistic performance. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 113–126). Brussels, Belgium: KVAB.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, *17*(1), 48–60.

Lee, H., & Oh, M. (2007). The effects of pre-task planning on L2 narrative tasks. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 127–138). Brussels, Belgium: KVAB.

Lee, H., Joo, K., Moon, J., & Hong, Y. (2007). EFL learner performance variation as the effect of interlocutor type. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 127–138). Brussels, Belgium: KVAB.

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

Lin, Y.-H., & Hedgcock, J. (1996). Negative feedback incorporation among high-proficiency and low-proficiency Chinese-speaking learners of Spanish. *Language-Learning*, *46*(4), 567–611.

Loban, W. (1966). *The language of elementary school children* (Research report no. 1). Champaign, IL: National Council of Teachers of English.

Nas, G. (1975). *Determining the communicative value of written discourse produced by L2 learners*. Utrecht, the Netherlands: Institute of Applied Linguistics.

Norris, J. M., & Ortega, L. (2003). Defining and measuring SLA. In C. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 716–761). London, UK: Blackwell.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601.

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47*, 101–143.

Roquet-Pugès, H., & Pérez-Vidal, C. (2007). Bilingual reading: An essential factor for the acquisition of written competence in a third language. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 183–192). Brussels, Belgium: KVAB.

Sempé, J.-J. (1973). *Rien n'est simple*. Paris, France: Editions Flammarion.

Skehan, P., & Foster, P. (2007). Complexity, accuracy, fluency and lexis in task-based performance: A meta-analysis of the Ealing research. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning & teaching* (pp. 207–226). Brussels, Belgium: KVAB.

Storch, N., & Wigglesworth, G. (2007). Writing tasks: Comparing individual and collaborative writing. In M. del Pilar Garcia Mayo (Ed.), *Investigating tasks in formal language settings* (pp. 157–177). Clevedon, UK: Multilingual Matters.

Storch, N., & Wigglesworth, G. (2010a). Learners' processing, uptake and retention of corrective feedback on writing: Case studies. *Studies in Second Language Acquisition, 32*(2), 303–334.

Storch, N., & Wigglesworth, G. (2010b). Students' engagement with feedback on writing: The role of learner agency/beliefs. In R. Batstone (Ed.), *Sociocognitive perspectives on language use and language learning* (pp. 166–185). Oxford, UK: Oxford University Press.

Tonkyn, A. (2007). Short-term changes in complexity, accuracy and fluency: Developing progress-sensitive proficiency measures. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 263–284). Brussels, Belgium: KVAB.

Van Daele, S., Housen, A., & Pierrard, M. (2007). Fluency, accuracy and complexity in the manifestation and development of two second languages. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 310–316). Brussels, Belgium: KVAB.

Van Daele, S., Housen, A., Kuiken, F., Pierrard, M., & Vedder, I. (Eds.). (2007). *Complexity, accuracy and fluency in second language use, learning & teaching*. Brussels, Belgium: KVAB.

Wigglesworth, G., & Storch, N. (2009). Pair versus individual writing: Effects on fluency, complexity and accuracy. *Language Testing*, *26*(3), 445–466.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998) *Second language development in writing: Measures of fluency, accuracy, & complexity*. Honolulu, HI: University of Hawaii Press.

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

APPENDIX A. ORAL DATA

The student is describing a cartoon strip (Sempé, 1973). It shows a lonely man who goes out to buy Christmas decorations and champagne. Once home, he decorates his apartment, changes into smart clothes, and opens the bottle. The champagne shoots out, the cork lodges in the ceiling decorations. As the man pours out the tiny amount of champagne remaining, the cork falls and smashes the glass.

The transcript has been divided into clauses (Foster et al., 2000) and, for the purposes of this illustration, has been pruned of hesitations, repetitions, and false starts. Each clause is scored according to the WCR, and the accuracy rating has been calculated:

	Clause	Category	Score
1	It's man	1	0.8
2	who is very lonely	accurate	1.0
3	and he decides	accurate	1.0
4	to go to shop	1	0.8
5	and to buy Christmas decorations	accurate	1.0
6	He is in the shop	accurate	1.0
7	And she buys them	1	0.8
8	And now he is going to his house	accurate	1.0
9	and decorate it	1	0.8
10	He prepare the table and the chair, the wine and the glass	1	0.8
11	And he opening the bottle	1	0.8
12	And when he open the bottle	1	0.8
13	the cork is going out to the wall	3	0.1
14	And when he putting out of for the glass	3	0.1
15	there is no wine in the bottle	accurate	1.0
16	And he surprising	2	0.5
17	because the cork go down	2	0.5
18	and smash the glass	1	0.8

Accuracy rating: $13.6 \div 18 = .76$

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press

APPENDIX B. WRITTEN DATA

These data are from an intermediate-level student essay on how education shapes future success. They have been divided into clauses (Foster et al., 2000). Each clause is scored according to the WCR, and its accuracy rating has been calculated:

	Clause	Category	Score
1	As we all know	accurate	1.0
2	exams are an important part of education in almost all schools and universities around the world	accurate	1.0
3	However, recently, there are some kinds of widespread concerns over the issue of the role of exams in education.	1	0.8
4	There are indeed a lot of advantages	accurate	1.0
5	that exams bring to education systems and students.	accurate	1.0
6	Firstly, exams can give the teachers chance	1	0.8
7	to evaluate	accurate	1.0
8	how their students are doing with their work	accurate	1.0
9	For most time, the test paper can reflect	1	0.8
10	how much knowledge the students have obtained	accurate	1.0
11	The second advantage of the exam system lies in its function	accurate	1.0
12	of compelling students	accurate	1.0
13	to study harder.	accurate	1.0
14	As we can see nowadays	accurate	1.0
15	in many schools and universities students study harder before exams	accurate	1.0
16	especially those who seldom sit down in the library in the beginning of a semester	1	0.8
17	are compelled	1	0.8
18	to leave their PC games and footballs	1	0.8
19	and begin	accurate	1.0
20	to study hard just before exams	accurate	1.0
21	However, there are a lot of disadvantages of exams	1	0.8
22	Firstly, the exam systems often make students	1	0.8
23	study just for the exam	accurate	1.0
24	Some students are even misled by the exams	1	0.8
25	by thinking	accurate	1.0
26	that getting a good score	1	0.8
27	is their aim of study	1	0.8
28	They try	accurate	1.0
29	to remember the knowledge in their book	1	0.8
30	Other time than for get their own idea of	2	0.5
31	how to put the knowledge into practice	1	0.8
32	For, as we all know	accurate	1.0
33	some exams aim at	accurate	1.0
34	testing the memory of students	accurate	1.0
35	Another disadvantage of exams comes with the cheating behavior of students during exams	accurate	1.0
36	Under the pressure of their parents and some rules of school some students resort	1	0.8
37	to cheating on exams	accurate	1.0
38	to get better score	1	0.8
39	According to my opinion exams are necessary in education systems	1	0.8

Accuracy rating: $35.3 \div 39 = .90$.

https://doi.org/10.1017/S0267190515000082 Published online by Cambridge University Press