Studies in Second Language Acquisition (2026), 1–20
doi:10.1017/S0272263126101892

METHODS FORUM

Exploring new measures of accuracy in second language task-based spoken performance

Peter Skehan[1] and Zhan Wang[2]

[1]Department of Culture, Communication and Media, University College London Institute of Education, London, UK and [2]Faculty of Education, The Education University of Hong Kong, Hong Kong, Hong Kong SAR
Corresponding author: Peter Skehan; Email: peterskehan@gmail.com

(Received 03 October 2025 UTC; Revised 29 May 2026 UTC; Accepted 22 June 2026 UTC)

Abstract

This article focuses upon the measurement of accuracy in second-language task-based spoken performance, with measures related to unit of analysis, error gravity, and accuracy linked to clause length. Building on established indices, three new measures of accuracy, all linked to clause length, are described. In addition, variations in an error gravity formula are also used. The measures, both established and new, are trialed with an existing dataset from a study exploring planning effects on second language narrative performance. A range of established measures is shown to generate appreciable effect sizes with this dataset. The newly developed measures also work well, as do alternative error gravity formulae. The new length-accuracy measures also seem sufficiently distinct from measures of language complexity. We discuss the implications of these findings for measuring accuracy in task-based spoken language performance, as well as the relevance of the new measures for theoretical accounts of second language performance.

Keywords: Error-free clauses; measures of accuracy; second language performance; second language speaking; task-based language teaching

In recent decades, a large literature has developed exploring the influence of task characteristics and task conditions on second language performance (Sasayama, Malicka, & Norris, 2025). There has also been associated theorizing, attempting to link such influences with performance, particularly Complexity, Accuracy, Lexis, and Fluency (CALF). Robinson's (2022) Cognition Hypothesis, for example, makes predictions regarding the impact of task complexity on language complexity and accuracy, while Skehan's (2018) Limited Attention Capacity approach suggests that, under some circumstances, there can be competition between different CALF areas.

Central to this literature has been the way performance itself is measured. As mentioned above, one significant response to the measurement challenge has been to assess task performance through CALF indices. The use of such indices is pervasive, and

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

as a result, a body of knowledge has accumulated, in benchmark fashion, against which new studies can be compared. In general, one can argue that such CALF-linked measurement approaches have not seen a great deal of development in recent years. Although there are alternatives in each area, many studies rely on the same set of measures: Clauses-per-AS unit (i.e. Subordination, for complexity), Error-free Clauses (EFC, for accuracy, Foster & Skehan, 1996), lexical density (for lexis, Malvern et al., 2004), and speed and breakdown (for fluency, Segalowitz, 2010). Alternative measures exist in each area, such as Words-per-Clause for complexity (Norris & Ortega, 2009), different accuracy measures (and see below), lexical sophistication and variation for lexis (Skehan, 2009), length-of-run and repair for fluency (Tavakoli & Wright, 2020). In most of these cases, the alternative measures do not correlate highly with the "standard" measures, so in all cases there is scope to explore these alternatives, uncovering sub-dimensions in each area, the links between different measures, and even optimal indices, a key challenge for the field. Here, we will mainly focus on accuracy and, to a lesser extent, complexity, and consider whether there is potential for improvement in the measurement options available.

Measuring accuracy

Recent research

Skehan et al. (2024) have published research that begins to suggest some of the alternative ways of measuring accuracy. They propose a distinction between measures which focus on unit, measures which link accuracy to length-of-clause (Skehan & Foster, 2005), and, following Foster and Wigglesworth (2016), measures which take the gravity of error into account. The first of these includes measures such as the proportion of error-free clauses (the commonest measure used in the field), or errors per AS-unit (O'Grady, 2023), or errors-per-100 words (Mehnert, 1998). In the second case, the focus is on how long a clause can be handled without likelihood of error. For the third, Foster and Wigglesworth (2016) argue that one needs to distinguish between different levels of error gravity. They suggest Level 1, 2, and 3 error levels, all defined in functional rather than morphosyntactic terms, and based on increasing degree of difficulty for the listener in recovering meaning (and see below).

Skehan et al. (2024) draw on three existing datasets (Bui, 2014; Skehan & Shum, 2014; Wang & Skehan, 2014), each aligned with a distinct theoretical focus: (1) a Conceptualizer-Formulator contrast (Levelt, 1989; Skehan & Shum, 2014), (2) time perspective (Robinson, 2022; Wang & Skehan, 2014); and (3) familiarity and pre-task planning (Bui, 2014). The study explored whether the new accuracy measures can illuminate more distinct or stronger relationships with the experimental variables, and indeed, Skehan et al. (2024) report that the measures make interesting but sometimes contrasting contributions. First, error-free clauses (EFC) performed generally well, but did show some inconsistency across studies, while the errors-per-100 words measures were more generally effective. Second, in the time perspective study, perhaps surprisingly, Level 1 (minor) and Level 2 (more serious) errors showed intriguingly different relationships with experimental variables, with more Level 1 errors in the here-and-now condition, but more Level 2 errors in the there-and-then condition (and see below for clarification with these error levels). Finally, there were issues with some length-linked measures, termed Length Accuracy 70% and Accuracy Lambda (see below), regarding unreliability with the first of these, and possible lack of construct uniqueness with the second, since it correlated highly with the Words-per-Clause measure of complexity (Skehan, 2018).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

We can conclude that these results are promising and support the value of exploring new accuracy measures. However, improvements are clearly needed, particularly with the need to identify effective length-linked accuracy measures that are distinct from complexity measures. Exploring some of these alternatives will be a major focus of the current research.

Constructs and measurement

We have noted that accuracy measures may have focus on unit, length, and gravity. Unit raises the question as to whether accuracy measures are structure-linked or not. EFC, and the corresponding errors per AS-unit measure, as well as potential weighted measures, such as Foster and Wigglesworth's (2016), are indeed structure-linked, and there is a potential linkage with a clause-based "unit of speech" discussion. In contrast, errors-per-100 words explicitly avoids any such notion, by-passing structural organization with the aim of gathering a more neutral sample of speech, characterized only in terms of the number of words used.

Length-accuracy, in contrast, originates from the practical need to avoid favoring speakers who use short, undemanding clauses, leading to higher but arguably inflated accuracy scores. However, there is the more serious issue that greater length may be more demanding of working memory while language is being produced, as it becomes important to keep track of connections between phrases and between clauses. As a result, there are possibly different lengths of clause, which function as a sort of threshold for different second language speakers. The remaining accuracy feature, gravity, allows us to give some credit to speakers who are "almost there." We also saw, in Skehan et al.'s (2024) re-analysis, that it can be insightful with some experimental variables, such as time perspective, where error gravity was related interestingly to the time perspective variable. Consequently, gravity-linked measures also have potential contributions to make.

Next, we need to try to relate the conceptual discussion to actual measures. The most widely used measure in the field, EFC, is strongly unit-based, with no concern for length or gravity. Mehnert's (1998) proposals for errors-per-100 words avoid any form of structural unit deliberately, for practical, sampling reasons. Foster and Wigglesworth's (2016) proposals effectively retain the clause-basis for measurement, but "layer" a gravity perspective on top. They propose a weighting for each clause, based on its error gravity: an error-free clause is scored as 1.0, a clause with level 1 error is scored as 0.8, a clause with level 2 error is scored as 0.5, and a clause with level 3 error is scored as 0.1. This can be represented as a 1-8-5-1 approach to scoring. The mean score can then be computed for each participant, giving an indication of overall weighted accuracy. Level 1 clauses have only minor errors that do not compromise meaning. For example, from the present data, Level 1 examples are: "so it's wakes him up again," and "to help it reading," rather than "read" in this second example. Level 2 reflects clauses that contain serious errors, but where meaning is recoverable, although not always obvious or quick. Examples are "It's noise him" (rather than "It disturbs him"), and "So just wave the curtain to one side," rather than "move," with this error being considered for Level 1, but then assigned Level 2 because "wave" is rather misleading and requires a little processing. Level 3 is characterized by very serious errors that make the intended meaning far from obvious, and possibly only partly recoverable. For example, "A hardmentship there are in the picture," is somewhat mystifying. (Further examples are provided below.) Clearly, using alternative weights to 1-8-5-1 is also possible, and would represent different hypotheses about the importance of particular levels of error gravity. We pursue this below.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

Exploring the link between accuracy and length is more complex. The two attempts from earlier research, Length Accuracy 70% and Accuracy Lambda, each have problems, as indicated above. The Length Accuracy 70% measure, in an ideal world, deals with data with clear patterns, and which shows a distinct 70% threshold, for example, with a sequence such as 2-word clauses (100% accuracy), 3-word (85%), 4-word (70%), 5-word (60%), 6-word (40%), and then lower accuracy still for longer clauses. This would lead to a score of "4": clauses longer than 4 words do not meet the 70% criterion, while all those of 4 words or fewer do. In practice, much data does not follow this ideal of a clear cut-off point, with lapses and inconsistencies across levels which require more subjective assessments of "correct" maximum length, introducing possible subjectivity and unreliability. In contrast, the Accuracy Lambda measure is not unreliable—a straightforward formula is applied. But, as noted above, it does correlate highly with the Words-per-Clause measure of complexity, suggesting some degree of conflation of complexity and accuracy constructs.

New measures

In view of this, some new length measures will be explored in the present study, which try to avoid a lack of construct uniqueness and unwanted correlations. All of them are clause-based, and they will be outlined briefly in this section, but with more detailed exemplification of calculations given in Appendix One. The first two, potentially, represent families of measures, with a particular implementation in the present study.

Error-free greater than (EFC>)

This group is similar to the basic EFC measure but is calculated only considering clauses above a certain length in words, such as EFC>4. The intention is to avoid inflated EFC scores by removing short correct clauses from the calculation. In pilot work, values from 3 to 6 were tried, and EFC>4 was chosen as most likely to produce significant results, and also a reasonable separation between particularly short clauses and those which would need to have some internal structure, with multi-phrase elements, and memory demands that would result from the need to relate morphosyntax across such elements, as in the case of, for example, agreement.

Proportion greater than (Prop>)

Such measures too are based on identifying clauses above a certain length in words. But in this case, with each qualifying word length, it is the average accuracy proportion, weighted for clause length, which is calculated. The proportion correct for each particular clause length is likely to decline as the clauses become longer, but this value is compensated for by the multiple based on clause length, representing a different way to reward a combination of length and accuracy.

WeightedNetAccuracy (WNA)

This is a revised version of the Accuracy Lambda measure, and the revision is aimed at reducing the correlation with the Words-per-Clause complexity measure. All the three-word clauses, four-word clauses, etc., are gathered together. Then it is the *net* number of clauses that is calculated for each clause length. In other words, one calculates, for each length, the number of correct clauses *minus* the number of incorrect clauses. This net result is multiplied by the particular length of the clause, and these terms are summed

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

and averaged. The final result may, logically, and indeed, often in practice, be negative, and as a result, this approach becomes a more stringent measure to reflect length-accuracy, since it takes account of all clauses for each length, not simply the correct ones. A worked example is provided in Appendix One.

The context for the present research

Next, we need to return to a point made earlier—that no new data will be reported here, but that the focus will be on how new measures can be more revealing with existing data. The original research, a Ph.D. study by the second author, investigated pre-task and online planning (Wang, 2009, 2014). In the context of a video narrative retelling, Wang (2009, 2014) explored four planning conditions: a *Watched group*, who had the opportunity to watch a video silently, and then re-watch and simultaneously retell the narrative; an *Online planning group* who were given more time to retell the video narrative (and see the details in Table 1); a *Watched online group*—effectively combining the two previous conditions; and a *Watched strategic group*, who did not have extra time during processing but who watched the video (without needing to speak), and were then given time to plan what they would say before re-watching the video and simultaneously describing it. A *Control group* was also involved. In the control condition, the learners started to narrate the story as soon as the video started to play; they did not have any pre-watching opportunity, or reduced time pressure from slowed video playback.

Wang (2014) used an EFC measure for accuracy and reported significant effects, consistent with the predictions that were made regarding the interplay of the different planning conditions (and see below). However, the study was limited if judged in relation to the range of accuracy variables that have since been developed. As noted earlier, one previous example of a comparable re-analysis is Skehan et al. (2024), which used some of the newly developed measures, although without the emphasis on length outlined here. However, that re-analysis was based on source studies which did not particularly emphasize planning, except as one of the experimental variables in one of the three datasets. As a result, the Wang (2014) dataset contains considerable potential to use the more newly developed measures to try to detect a wider range of planning effects. More specifically, the dataset enables us to probe how pre-task and online planning can have an impact on different sub-dimensions of accuracy, and whether, with such variables, more nuanced measurement can be revealing, and demonstrate relationships which did not emerge in the original work.

Table 1. Details of the research design

	Speaking conditions	Pre-watching (5 min)	Online planning (8 min speaking time)	Strategic planning (3 min speaking time)	Original video (5 min speaking time)	Task time (min)
1	Control	–	–	–	√	5
2	Watched	√	–	–	√	10
3	Online planning	–	√	–	–	8
4	Watched online planning	√	√	–	–	13
5	Watched strategic planning	√	–	√	√	13

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

Practical issues

Clarification is needed on several of the decisions concerning the coding methods that were used. First, for reasons of practicality and time-demands, fairly simple methods were used with the basic unit of words, which were defined as what occurred between spaces. As a result, we did not code formulaic or multi-word expressions differently from other words. In addition, contractions were expanded, and so "isn't," for example, would be treated as "is not." These decisions were motivated by practicality.

Second, we need to clarify how we treated error. Essentially, we followed Foster and Wigglesworth's (2016) functional approach, and the important point is that error is judged through impact on communication, and the comprehension effort required to understand what was said. The levels of error we used (1,2,3) should be viewed through this lens. Accordingly, while morphosyntactic and lexical errors were important, and perhaps the commonest basis for error codings, phonological and pragmatic factors could also figure, if they had an impact on comprehension.

This connects with the third point, error gravity. Given this emphasis on functional impact, rather than any scale of morphosyntactic seriousness, the criterion for the three levels of error was how much comprehension effort was required. Level 1 is minimal, exemplified by a third-person -s slip—noticed perhaps, but not disruptive. Level 2 errors were those that led to a brief pause in decision-making when the coder required a moment to check that meaning had been correctly extracted. For example, (and with a Level 1 error also shown, together with pausing (bracketed, in seconds)): *Oh (0.43) God, he use the toy pistol (error level 1) to clean his wash (error level 2).* Level 3 errors were those where it was either very difficult to recover meaning, or even impossible, as with: *Oh suddenly (1.32) and performance is changed to the same (error level 3).* Context did not allow any insight as to what is meant in this case.

As indicated above, we explored a few different weightings for error. These were 1-8-5-0, a slight modification of Foster and Wigglesworth's (2016) original 1-8-5-1 weighting, and also 1-5-2-0 and 1-3-0-0. None of the weightings we used gave any credit for clauses with Level 3 errors: we decided that if meaning is impossible or close to impossible to recover, no credit should be given. In addition, the new weightings for Level 1 and Level 2 errors were less generous than those used in Foster and Wigglesworth (2016). This was less motivated by principle, as was the case with the Level 3 weighting of zero, and more concerned with exploring the strength of experimental effects when less credit was given for the two intermediate types of error. The Level 1 error weightings that we used were 0.8, 0.5, and 0.3, suggesting, respectively, a weighting of almost as good as no error (0.8), the half-weighting (0.5), which gave Level 1 errors distinctly lower but still appreciable value, and the final Level 1 weighting of 0.3, which effectively gave little credit. These three values seemed to span the range of weightings fairly comprehensively and systematically. On a similar basis, the three Level 2 weightings of 0.5, 0.2, and 0, reflecting the seriousness of Level 2 errors, move from quite a generous weighting for what is an error that impairs communication, to values that give little or no credit. Essentially, all values were exploratory at this stage, intended to sample the possibilities available.[1]

[1]The computer program that calculated the error gravity measures output many more scores than those reported here, with more variation in how Level 1 and Level 2 errors were handled. Scores ranged from 1-8-5-0, through 1-8-4-0, 1-8-3-0, and so on down to 1-3-1-0, and 1-3-0-0, with many combinations in between. It is our intention to explore in the future whether there are any generalizations in the effects that are found with particular combinations of weights. We also hope to do this with other datasets. For now, the values we chose

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

Research questions

1. Do length-linked accuracy measures correlate with measures of structural complexity, particularly the Words per Clause measure?
2. What is the impact of the experimental conditions, contrasting a range of watched, online and strategic planning conditions, on accuracy and complexity measures, relative to the control group?
3. Do any of the experimental conditions differ significantly from one another?
4. Which measures of accuracy are most effective, as judged through effect sizes? Within this general question, and of particular and more specific interest, which accuracy measures reflect greater impact on performance than the EFC measure?

The first question focuses on a problem identified in the original study regarding the relationship between length accuracy measures and the Words-per-Clause complexity measure. The second and third extend the original study to include the newly developed measures. Then the fourth question is concerned with the central issue of whether there is any added value from the new measures that have been developed, either practically for future research or theoretically.

Methodology

Participants

The participants (N = 87) in this study were undergraduates at a university in Hong Kong, following different majors, and aged between 18 and 22. All were native Chinese speakers, with Mandarin as their L1, who had learned English as a second language. None had been overseas for longer than three months. They were recruited on a voluntary basis and received a small fee for participation in the study. Data collection was on a one-to-one basis. Participants first completed an English proficiency test (see below), which was then used to assign individuals to experimental conditions. This assignment balanced groups for proficiency test scores, gender, university year, and major. Participants' English proficiency, from self-report, ranged from TOEFL 540 to 630, and IELTS 6 to 7.5.

English proficiency test

The pre-test used a version of the TOEFL Listening subtest, extracted from Hinkel (2004). This type of test was used because it has been reported to be a strong correlate of the general English proficiency measured by TOEFL examinations (Sawaki et al., 2009).

Task and time operationalization

A Mr. Bean video was used to elicit narrative-based speaking performance. The story it portrays is that Mr. Bean is having trouble getting to sleep one night. He resorts to a variety of solutions to this problem, finally succeeding after a range of rather bizarre activities. The same video was used in each condition. From previous research (Skehan & Foster, 1999), it has been established that Mr. Bean videos are appropriate for narrative

in the present study, as indicated in the main text, are simply to cover this span of possibilities in a representative manner, on an exploratory basis.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

retelling because each video is short, largely without explicit language content, easy to comprehend, and appealing. The current video had been trialed with comparable participants to ensure that it was acceptable, comprehensible, and engaging.

Earlier, it was argued that operationalizations of picture-based online planning conditions have lacked standardization because of the uncontrolled handling of time pressure. To avoid this problem, the present study uses video-based narratives, and this was the foundation for increasing standardization. This rationale is in line with the online planning definition—"moment by moment preparation during speech" (Ellis & Yuan, 2005), and careful control of processing during speech production.

The need, therefore, for the two online conditions in the study was to find a way to slow down the speed of presentation of the video, which did not render the input unnatural. Piloting with different slowed rates was carried out with comparable participants, who were monitored and questioned regarding their response to the different videos. Three versions were tried, at 50%, 60%, and 75% of normal speed, using Adobe Premiere video compilation software. Based on trial participants' feedback, the 60% version was used, making an original 5-minute video last 8 minutes. This was regarded as natural in appearance, and it therefore gave controlled additional time for online planning to occur.

Research design, task conditions, and instructions

We compare five different conditions in this study, as shown in Table 1. The table clarifies the amount of time involved in each study, and how pre-task and online planning conditions were used.

The "Control" condition (Condition 1) gives no planning and no pre-watching opportunities for speakers. In its operationalization, a speaker was sitting in front of a computer. The researcher asked the speaker to narrate while watching a normally played video. The speaker had no opportunity for pre-watching the video or knowing the story.

The "Watched" condition (Condition 2) allows speakers to watch the video before narrating the story. This manipulation is intended to reduce time pressure at speech conceptualization by exposing the story content to speakers first. In this way, more attention is available to attend to formulation while the speaker is performing the task.

The "Online Planning" condition (Condition 3) has no pre-watching opportunity, but it provides extra time for speakers to conduct online planning with the slowed video. It is hypothesized that the extra time for conducting online planning can help learners attend to morpho-syntactic formulation so as to increase speech accuracy.

The "Watched Online Planning" condition (Condition 4) provides the opportunity for pre-watching, as well as the extra time for speakers to conduct online planning while narrating. It is likely that having the chance to watch a video before speaking *and* then benefitting from additional time *while* speaking will reduce time pressure at both the conceptualization and the formulation phases so that learners can speak with increased accuracy and complexity (Ellis & Yuan, 2005).

Finally, the "Watched Strategic Planning" condition (Condition 5) provides an opportunity for pre-watching *and* extra time for speakers to conduct strategic planning before speaking. It reinforces the "Watched" condition (Condition 2) by providing extra time before narrating, so learners can use the time either for content preparation or problem anticipation (see Wang, 2014). Therefore, this condition could have a similar effect to task conditions involving adequate pre-task planning time (Mehnert, 1998; Ortega, 1999; Skehan & Foster, 1997).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

Measures of spoken performance

The measures used in the present study are shown in Table 2. Most measures are self-explanatory, but clarification is needed that there are three versions of the error gravity measures, using different combinations of weightings to capture contrasting interpretations as to how to deal with error gravity, and the importance of Levels 1 and 2 errors. Following Skehan et al. (2024), the errors-per-100 words measures separate the minor (Level 1) and serious errors (Level 2). In addition, there are five measures that reflect the length of a clause linked with accuracy, and example calculations for these are shown in Appendix One. Finally, to explore the accuracy-complexity relationship, two measures of complexity are included: subordination, the most widely used complexity measure, but also Words-per-Clause, the measure correlated with some length-linked accuracy measures (Skehan et al., 2024).

Coding

First, the data were coded following CHAT conventions (MacWhinney, 2025) to enable compatibility with CLAN programs. Second, the transcribed data were coded in ways more appropriate for second language task-based performance data. Each participant contribution was coded into AS units (Foster et al., 2000), then coded for accuracy, with

Table 2. Dependent variables used in the study

Variable	Score calculation method
Error-free clauses (EFC)	Number of error-free clauses divided by the total number of clauses
Error gravity 1-8-5-0 (Weighted 1-8-5-0)	An EFC measure calculated as 1 × number of correct clauses + 0.8 × number of clauses with only minor, Level 1, errors + 0.5 × number of clauses with serious (Level 2) errors
Error gravity 1-5-2-0 (Weighted 1-5-2-0)	The same approach but with multiples of 1, 0.5, and 0.2, respectively
Error gravity 1-3-0-0 (Weighted 1-3-0-0)	The same approach but with multiples of 1, 0.3, and 0
Errors-per-100 words, minor (E100 Minor)	Total number of Level 1 (minor) errors divided by the total number of pruned words
Errors-per-100 words, serious (E100 Serious)	Total number of Level 2 (serious) errors divided by the total number of pruned words
Error-free clauses longer than 4 words (EFC>4)	Another EFC measure but only based on the proportion of accurate clauses that are five words or greater in length
Proportion measure greater than 4 words (Prop>4)	A clause length weighted average proportion correct measure for clauses more than four words
WeightedNetAccuracy (WNA)	A weighted average measure of the net (correct minus incorrect) clause accuracy
Accuracy Lambda (AccLambda)	The (raw) number of error-free clauses of each length is multiplied by the length of the respective clause. These are summed and divided by the total number of clauses
LengthAccuracy 70% (LenAcc70)	The maximum length of clause that is used at 70% accuracy or more
Subordination	The total number of clauses divided by the total number of AS units
Words per clause	The average number of words that is used per clause

Note: Calculated examples for length-accuracy measures are shown in Appendix One. Descriptions in parentheses are used in the tables that follow.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

codings showing not simply error but also error gravity (Foster & Wigglesworth, 2016). Next, data was coded for features of fluency, including speed, pauses, and repair. This data was used to calculate the number of pruned words, relevant for some accuracy measures. Coding was primarily conducted by a research assistant with an MA degree in Language Education. To establish intercoder reliability, the second author independently coded 10% of the data. The results showed inter-rater coding correlation coefficients of 0.92 for clause and AS-unit codings, and 0.87 for a composite EFC score. The more specific inter-rater coefficients were 0.84 for Level 1 errors and 0.71 for Level 2 errors. Although the Level 2 error figure is slightly low, we consider it sufficient for our current research purposes. One factor was that there were far fewer Level 2 errors, and so the variation was not so great, which possibly attenuated the correlation that was found. All discrepancies were resolved through discussion to ensure coding consistency. Third, the data were analyzed using TaskProfile (Skehan, 2014a), a computer program that takes coded data and generates a very large number of indices of complexity, accuracy, fluency, and lexis.

Analysis

The present study is exploratory in nature. As a result, relatively simple inferential statistics are used. The primary analytic approach was a One-way Analysis of Variance, followed by Bonferroni-adjusted t-test results, comparing the different experimental conditions with the control group. These results provide an indication of the strength of the relationship through the level of significance achieved. But more importantly, effect sizes were calculated. The sample sizes in the study were not large, and as a result, significance was difficult to achieve, given the low statistical power. As a result, the effect size values perhaps provide a clearer indication of where relationships were strong, and underpin most of the argumentation. In all cases, the effect sizes are Cohen's D.

All calculations were carried out using R (R Core Team 2022). The ANOVA and t-test calculations are based on the *psych* package, and the effect sizes on the *lsr* package.

Results

The descriptive statistics (mean, standard deviation) across the experimental conditions are shown in Table 3.

A series of generalizations follows from these figures. First, for every measure, the Control group produces the lowest scores (or with Errors-per-100 words, the highest). Second, the Online condition is associated with most of the highest scores. The Watched Online condition does not produce the highest score on any of the measures. The Watched condition produces the highest score for the Length Accuracy 70% score and also the Words per Clause score. The Watched Strategic condition produces the lowest (i.e. best) Errors-per-100 words (Level 1) error, as well as the highest Accuracy Lambda score.

Earlier, we discussed the high correlation between Accuracy Lambda and Words per Clause. We included the Accuracy Lambda measure in this re-analysis to see whether this high correlation would be repeated. The correlation is 0.88, highly relevant to Research Question 1. Once again, we have to draw the conclusion that this putative measure of accuracy is, in fact, very close to a measure of one facet of complexity. Accordingly, we will not discuss it further.

Table 4 provides inferential statistics on the different measures, directly addressing Research Question 2. The significance level is given from a one-way analysis of variance

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

Table 3. Descriptive statistics

	Control		Watched		Online		Watched online		Watched strategic	
	Mean	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*
EFC	.47	.10	.57	.12	.65	.13	.62	.14	.61	.11
Weighted 1-8-5-0	.87	.03	.90	.03	.92	.03	.91	.04	.91	.03
Weighted 1-5-2-0	.71	.06	.77	.07	.82	.07	.80	.08	.79	.06
Weighted 1-3-0-0	.61	.08	.69	.09	.75	.09	.72	.10	.72	.08
E100 minor	10.50	2.6	8.94	3.30	7.13	2.82	7.71	3.66	7.11	2.58
E100 serious	1.51	.98	.80	.70	.47	.41	.80	.68	.93	.66
EFC>4	.50	.12	.58	.12	.65	.14	.62	.14	.64	.12
Prop>4	1.44	.56	2.68	1.72	3.00	1.06	2.56	.88	2.64	1.06
WNA	1.20	1.10	.65	1.53	1.41	1.49	1.14	1.62	1.20	1.39
LenAcc70	9.31	1.89	10.8	1.52	10.7	1.08	10.7	1.08	10.1	1.09
AccuracyLam.	4.70	.63	5.17	.77	5.09	.43	5.12	.53	5.20	.47
Subordination	1.34	.11	1.40	.13	1.47	.19	1.37	.14	1.39	.16
Words per clause	5.04	.61	5.43	.78	5.28	.48	5.34	.44	5.34	.46

Note: Table 2 provides a detailed description of the measures (dependent variables) reported in this table. See also Appendix One for further details.

Table 4. One-Way ANOVA, T-Test, and effect size results

	Oneway ANOVA	Watched Sig.	Eff.Siz	Online Sig.	Eff.Siz.	Watched online Sig.	Eff.Siz.	Watched strategic Sig.	Eff.Siz.
Error-free clauses	.002	.43	.90	.003	1.48	.03	1.21	.04	1.35
Weighted 1-8-5-0	.003	.09	1.02	.001	1.66	.009	1.24	.01	1.29
Weighted 1-5-2-0	.002	.21	.98	.001	1.60	.02	1.24	.02	1.36
Weighted 1-3-0-0	.002	.30	.96	.002	1.57	.02	1.23	.02	1.37
Minor errors	.01	.98	.52	.06	1.24	.25	.86	.05	1.31
Serious errors	.01	.08	.85	.001	1.44	.08	.86	.32	.71
Error free > 4	.02	.90	.51	.03	1.27	.28	.93	.18	1.05
Proport. > 4	.0001	.04	.93	.003	1.79	.10	1.49	.06	1.37
WNA	.002	.44	.88	.007	1.46	.04	1.18	.03	1.37
LenAcc70	.11	.05	.89	.10	.92	.10	.92	.90	.55
Subordination	.27	.90	.50	.34	.79	.93	.17	.90	.29
Words per clause	.56	.77	.57	.90	.46	.90	.59	.90	.57

Note: Table 2 provides a detailed description of the measures reported in this table, and example calculations are provided in Appendix One.

for all five groups. Then, each experimental condition is compared with the control group performance. For each, the t-test significance is shown (with Bonferroni correction), and also the effect size. Given that sample sizes were so small, and statistical power low, the effect size associated, generally, with reaching the .05 significance level is quite high, at around 0.89.

The three measures that do not show significant ANOVA results, LenAcc70, Subordination, and Words per Clause, are connected with length and complexity. We will return to the Length Accuracy 70% measure, but will not discuss the other two measures further, given these results.

We can consider the remaining effect sizes in terms of the experimental conditions (Research Question 2) and the different measures of accuracy (Research Question 3). Regarding the first of these, we argue that there is something of a ranking, along the lines Online > Watched Online = Watched Strategic > Watched. In other words, the Online condition is associated with the largest accuracy effect sizes, on almost every

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

measure. Some of these effect sizes are very large indeed, across most measures. The Watched Online and Watched Strategic measures also generate large to very large effect sizes. Perhaps the Watched Strategic condition slightly outperforms the Watched Online condition, but with the Proportion measures, this is reversed. Finally, the Watched condition produces effect sizes which are more mixed, with some small to medium, but with others (EFC, the weighted measures, and the proportion measure) close to large. Consequently, we can claim that there are several ways in which performance can be enhanced, though pre-watching alone seems the least effective method.

With Research Question 4, here too, there are interesting differences. We will judge these figures, also, through effect sizes, and explore which measures have effect sizes greater than the baseline EFC result. Two types of measures stand out in this regard. Most clearly, the Prop>4 measure generates the strongest effects. This is proposed to measure capacity to operate with less error in clauses which are longer, as do the EFC>4 and WeightedNetAccuracy (WNA) measures. Prop>4 differs from EFC>4 since it incorporates a weighting for greater length. It differs from WNA since it does not penalize speakers who use long clauses incorrectly. In other words, it reflects length, but takes what might be termed a positive approach to scoring. The Prop>4 effect sizes range from near-large (0.93) up to very large (1.79), and all results exceed the corresponding EFC result. This measure shows its greatest difference with the "base" EFC measure, specifically for the two online conditions. Having less time pressure in speaking seems to link with the ability to produce more, longer, correct clauses[2].

Next, in order of strength and consistency, though only narrowly larger than the EFC effect sizes, are the weighted clause error measures, which produce average effect sizes of 1.30 (1-8-5-0), 1.30 (1-5-2-0), and 1.28 (1-3-0-0). Taking a less severe approach to error seems to capture slightly stronger effects than the standard EFC measure, as well as the two errors-per-100 words measures. Other measures do not perform as strongly. Even so, one such measure does merit brief attention. This is the WeightedNetAccuracy (WNA) measure, which slightly exceeds the EFC result for the Watched Strategic condition, and is very close indeed, though lower, for the other three conditions.

These results start to address the "value added" aspect of this study, in that measures other than EFC do seem to show greater or at least equal sensitivity to detecting differences between conditions. The differences with EFC may not be large, but particularly with small studies, such as this one, they may be very important. The actual effect sizes shown here are often very large, unusual with accuracy measures, as indicated by the average effect sizes reported by Sasayama et al. (2025), at 0.71 for pre-task planning and 0.80 for online planning.

The other major comparison in this study is between the different experimental conditions themselves (Research Questions 3 and 4), addressing the question of whether different preparation conditions have different impacts on performance. None of the detailed comparisons, i.e. those between the various experimental conditions, were significant after appropriate corrections. Bearing in mind, once again, the small samples, and the level of effect size needed to generate significance (at around 0.89), one

[2]A reviewer raised the issue of how much data was lost in the calculations of the ErrfrGr4 and Propgr4 measures. For clauses, this ranged from 37.4% to 45.7%, and for words, the corresponding figures are 23.5% and 31%. P values from a One-way analysis of variance in these cases were .36 and .37, suggesting no significant difference in data removal across conditions.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

can look at the largest experimental group difference—that between the Watched and Online conditions. Effect sizes here were low with the measures linked to length (WeightedNetAccuracy, Length Accuracy 70%, and the EFC> than measures). Clause-linked measures generated effect sizes in the .50 to .70 range, suggesting that, with larger sample sizes, significances might be found. But in the main, the range of new measures was not really associated here with more effective detection of significant differences between the experimental conditions.

Discussion

We will explore two main themes in this section. First, we discuss the measurement of accuracy in spoken language task-based performance, with a focus on comparing the different measures and their implications for future research. Second, we will look at the effects of the different experimental conditions in the study, and explore whether the new measures that have been used lead to any new insights in relation to planning effects.

Regarding the first theme, the previous literature has generally relied on the error-free clauses measure. As we have seen, alternative approaches have been proposed. All these innovations have been largely theory-neutral and motivated by practical measurement problems. Mehnert (1998) argued that an errors-per-100 words measure might work better with the specific L2 in her study (German). Skehan and Foster (2005) wanted to find ways of minimizing the impact of (correct) use of very short utterances. Foster and Wigglesworth (2016) point out that in an EFC analysis, minor errors have the same weight as major errors. We have tried to build upon these insights and also introduce a more conceptual perspective.

Based on the present study, we offer three main generalizations regarding measurement. First, the basic EFC measure, as well as the range of weighted scores (1-8-5-0, 1-5-2-0, 1-3-0-0), function well. The weighted measures do have a (very) slight edge with most experimental conditions, consistent with all three datasets in Skehan et al. (2024). Using such weighted ratios is therefore recommended for future researchers. Each different weighting reflects a slightly contrasting interpretation of the importance of different levels of error gravity, and further research with different weightings is desirable.

Second, the errors-per-100 measures also work well, but they are not quite as sensitive as some other measures. EFC and the weighted measures deliver effect sizes in the range 1.23 to 1.30, whereas the errors-per-100 measures are lower, with 0.97 (for Level 1 errors) and 0.98 (for Level 2 errors). These findings are broadly consistent with the results reported in Skehan et al. (2024). Even so, further research is needed to explore whether the absence of a very strong contrast here with the two errors-per-100 words measures (Levels 1 and 2), and its presence in the other datasets, is related to the different experimental variables that were used. The present study has more of a Formulator emphasis, whereas the studies in Skehan et al. (2024) were more Conceptualizer-oriented (Levelt, 1989).

Third, it is the length measures that are the most interesting, and where the value-added component of the present research is clearest. We will, though, start with the least effective measure, Length Accuracy 70%, whose ANOVA was not significant, with all effect sizes below 1.0. This measure has a moderately high correlation with the Words-per-Clause measure of complexity, at .58. Given that neither complexity measure, Subordination or Words per Clause, shows strong experimental effects, the performance of this accuracy measure is not surprising.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

In contrast, the other length-linked measures produce greater average effect sizes across the four conditions, at 1.04 for EFC>4, 1.22 for WeightedNetAccuracy, and 1.40 for the Prop>4. The first of these measures is essentially a version of the EFC measure—simply based on the elimination of shorter clauses, and judged as such, it produces lower effect sizes than the other EFC measures. The remaining length measures are different. They do have a clause basis, but each involves a method of reducing the impact of a confound with clause length. This is reflected in the lower correlations of these newer measures with the Words-per-Clause measure, at 0.29 for WeightedNetAccuracy and 0.48 for the Proportion measure—not a negligible correlation, but suggesting sufficient distinctness from complexity to justify discussion as an accuracy measure. The first of these two measures, WeightedNetAccuracy, attempts to neutralize clause length effects by counting the *net* number of correct clauses. In other words, using longer clauses incorrectly can incur a penalty. The second rewards length but on the basis of an *average* proportion, and in this case, there is no penalty for using longer clauses incorrectly—simply a lower score. These measures, then, are more successful not only in being more distinct from general clause length, but also in their effect sizes—the Prop> 4 measure produces the largest effect size in the entire study.

Taking the present results with those from Skehan et al. (2024), a broader picture of accuracy measurement is emerging, in terms of error-free clause measures, errors-per-100 words, and measures linked to length. The general pattern in the results confirms the utility of the first two of these. EFC, the workhorse measure in our field, is effective, and an additional error gravity variation is clearly worth pursuing. The edge it gives is small but consistent. Similarly, the errors-per-100 words measures are also largely confirmed as useful, even if they deliver slightly lower effect sizes. Clearly, though, the length-linked measures perform consistently well, especially Prop>4, and this is of some importance for future research in the field. This is the first study to report results with this measure, and it is clear that wider use is well worth exploring.

One can also examine these different accuracy measures in relation to the different experimental conditions. The average effect sizes, across conditions, and in increasing order, are Watched (0.91), Watched Online (1.17), Watched Strategic (1.25), and Online (1.55). In other words, all these planning conditions are very effective. Even the worst, the Watched condition, produces an average effect size which is medium-large in nature, and others are all large or very large. An interesting contrast within the experimental conditions is between pre-task and online planning. The three conditions, which do implicate a pre-task phase (Watched, Watched Strategic, and most interesting of all, Watched Online), while producing impressive average effect sizes, nonetheless produce lower average effect sizes than the (pure) Online condition (without any pre-task activity). Speaking with a slowed speed of input is what leads to the least error in this narrative task, even though there has been no preparation. Returning to the result for the Watched Online condition, one might have expected an additive influence here, capitalizing on the combination of preparation *and* the slowed speaking conditions, but it seems as if the Watched preparation conflicted in some way with the Online performance.

However, one additional insight suggests an interaction between planning conditions and accuracy measurement options. The outstanding performance of the Prop>4 measure, and also the weighted measures, is with the Online (especially) and Watched Online conditions. The effect sizes are very large in these measure-by-condition combinations, and so it seems that the Online condition is particularly advantaged with the use of these particular indices. What is interesting is that the two measures, Prop>4 and the Weighted measures (1-8-5-0, 1-5-2-0, 1-3-0-0), have in common some

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

level of generosity. The Prop>4 measure, as indicated earlier, is a "positive" measure—rewarding correct use of longer clauses but not penalizing incorrect use. The weighted measures give some credit for minor (especially) and even serious errors. Interestingly, in that respect, the "1-8-5-0" interpretation (the closest to the original proposed by Foster & Wigglesworth, 2016, and the most generous) produces the highest effect size for the two online conditions. Possibly, when there is less time pressure in performance, participants can access and try to use less controlled language, with more time for more effective lemma retrieval and formulation, as well as more attentional resources available for monitoring. There is a greater risk of error, but possibly error that does not prevent meaning from being expressed clearly enough to enable comprehension. Similarly, less time pressure can facilitate the use of longer and more complex clauses, but again, where there is a greater likelihood of error. The Prop>4 measure does not directly penalize error in such clauses. Lower scores result, but this is a more generous approach. If this post-hoc interpretation has any validity, it suggests some accuracy measures may be appropriate in giving credit to emerging language.

We can now relate these results to the nature of the speaking process (Kormos, 2006; Levelt, 1989; Skehan, 2022a). It appears that pre-task preparation is generally beneficial for accuracy, but that simply being exposed to material (Watched) has limited effectiveness. In contrast, being so exposed and then having the opportunity to process that material (Watched Strategic) confers an additional advantage. The advantages in each case seem to connect with the Conceptualizer stage and the development of the propositions to be expressed. Then the online condition potentially connects with two stages within the speaking process. More time available enables better retrieval of Conceptualizer traces (Wang, 2014), so that the pre-verbal message can be more fully realized. It also supports the major sub-processes of the Formulator stage, facilitating deeper and more extensive lemma retrieval, with associated syntax building. Given the limitations of the second language mental lexicon and the likelihood of serial processing (Kormos, 2006), this may be vital. Finally, less time pressure, particularly in such a serial processing context, can enable more effective and less disruptive monitoring.

Taken together, these possibilities may help us understand why more generous scoring raises the effect sizes with the weighted measures and leads to higher effect sizes with the length-linked Prop>4 measure. With the first of these, possibly a little more time enables greater access to emerging interlanguage, of the sort that is yet to be fully controlled. The time, in other words, allows access to less stable, more vulnerable aspects of language. We are dealing with error in language, but error that may have interesting connections to development. Regarding length-linked accuracy, possibly here there is more effective recall of previous Conceptualizer work, perhaps linked to more attention available for monitoring (Skehan, 2022b). Clearly, these are only speculations, but it is interesting that a CALF perspective on general performance might have some potential to explore more underlying psycholinguistic and developmental processes.

Implications

Two major implications emerge from the present research. First, we have argued that the different accuracy measures potentially tap slightly different, but related aspects of a general accuracy construct. However, as shown in Skehan et al. (2024) and also here, research studies vary, and there is no consistent "best" measure that might be a candidate to replace EFC. We argue that the clearest implication is that research studies should not rely on only one index, but should draw on a range of measures to try to relate the different measurement options to the emphases in each study. In detail, we

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

would suggest that a clause-based measure is essential, with associated weightings, as is an errors-per-100 words measure, separating levels of error gravity. Finally, some attempt to incorporate length accuracy is important. In general, the Prop>4 measure seems to be, on current data, the most impressive, in terms of consistency and also the strength of the effect sizes reported.

A second implication relates to the operationalization of an online planning condition. We argued earlier that there is potential non-standardization of an online planning condition because of individual differences in participants as they respond to demands linked to time pressure. In the present study, we used a video-based presentation for a narrative, and were able to control and modify the speed of presentation carefully. We would argue that this has a great deal to commend it, and indeed, would be worth using by others researching online planning.

More generally, and beyond the specific implication regarding the operationalization of online planning, the research suggests that time pressure has major effects on task-based spoken language performance. The present study has explicitly focused on this area, attempting to control time pressure very carefully. It is likely, though, that time pressure is relevant in many research designs, and may be a disguised but vital variable for a broad range of task-based studies of spoken language performance.

Limitations and future directions

There are a number of limitations to the present research. The numbers of participants, in specific sub-conditions, are clearly smaller than is ideal, leading to a reliance on effect sizes, rather than significances, for some of the argumentation. There is also the problem that only one task, a video-based narrative retelling, was involved. It would be desirable to explore other narrative tasks (different topics; picture-based as well as video-based), and indeed, other task types (such as interactive, or with personal-linkage) to gain a better idea of the generality of the claims which are made. Finally, the participant base was relatively narrow—only one L1 (Chinese), and drawn from a university-student population. Taking account of all these limitations in future research would strengthen any claims that might be made.

Beyond remedying these limitations, there are, though, other directions for future research. We will mention two of these. First, better measurement of accuracy can, potentially, link with theory more effectively. The Cognition Hypothesis, for example, makes predictions regarding the accuracy-complexity relationship. A viewpoint emphasizing Leveltian processes such as Conceptualization, Formulation, and Monitoring, and assuming attentional limitations, also presupposes effective accuracy measurement. The range of measures discussed here might enable a more nuanced view of these theoretical positions and enable predictions to be tested more effectively. Second, and essentially developing the last point, there is scope to explore whether different measures of accuracy might interact with different experimental variables. We saw how certain length-oriented measures were particularly effective with online planning. Skehan et al. (2024) report interesting linkages between error gravity and time perspective. There is considerable potential to extend this work, given the greater range of accuracy measures now being proposed.

Conclusion

The broad conclusion from the present study is that second language task-based research into speaking has relied, where accuracy measurement is concerned, on a narrow view of what can be done, and has been too dependent on the EFC measure.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

We have shown that there are alternative measures and that they link, meaningfully, with features such as unit of measurement, seriousness of error, and length of clause. In addition, the measures reported here, as well as in Skehan et al. (2024), have generated insights regarding task-based spoken language performance, which have extended our understanding of how experimental variables may exert an influence on performance. As a central implication, we suggest that research studies should be more varied and principled in their approach to accuracy measurement.

Acknowledgments. We would like to express our thanks to three SSLA reviewers, as well as the Handling Editor, for their comments on this article. Their comments have been extremely helpful and have strengthened and clarified the article considerably.

Competing interests. We have no known conflict of interest to disclose.

References

Bui, H. Y. G. (2014). Task readiness: Theoretical framework and empirical evidence from topic familiarity, strategic planning, and proficiency levels. In P. Skehan (Ed.), *Processing perspectives on task performance*. (pp. 63–93). https://doi.org/10.1075/tblt.5.03gav

Ellis, R., & Yuan, F. (2005). The effect of careful within-task planning on oral and written task performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 167–192). John Benjamins. https://doi.org/10.1075/lllt.11.11ell

Foster, P. & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition*, *18*, 299–324. https://doi.org/10.1017/s0272263100015047

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375. https://doi.org/10.1093/applin/21.3.354

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, *36*(1), 98–116. https://doi.org/10.1017/s0267190515000082

Hinkel, E. (2004). *TOEFL test strategies with practice tests* (3rd ed.). Barron's.

Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum. https://doi.org/10.4324/9780203763964

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511812034

MacWhinney, B. (2025). Understanding language through TalkBank. *Current Directions in Psychological Science*, *34*(2), 75–81. https://doi.org/10.1177/096372142400000001

Malvern, D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Palgrave Macmillan. https://doi.org/10.1057/9780230507947

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, *20*(1), 52–83. https://doi.org/10.1017/S0272263198001041

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555–578. https://doi.org/10.1093/applin/amp031

O'Grady, S. (2023). Halo effects in rating data: Assessing speech fluency. *Research Methods in Applied Linguistics*, *2*(2), 100048. https://doi.org/10.1016/j.rmal.2023.100048

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition*, *21*(1), 109–148. https://doi.org/10.1017/S0272263199001058

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Robinson, P. (2022). The cognition hypothesis, the triadic componential framework, and the SSARC model: An instructional design theory of pedagogic task sequencing. In M. J. Ahmadian & M. H. Long (Eds.), *The Cambridge handbook of task-based language teaching* (pp. 205–225). Cambridge University Press. https://doi.org/10.1017/9781108687308.010

Sasayama, S., Malicka, A., & Norris, J. M. (2025). *Second language cognitive task complexity: A research synthesis*. John Benjamins. https://doi.org/10.1075/tblt.18

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL internet-based test. *Language Testing*, *26*(1), 5–30. https://doi.org/10.1177/0265532208098164

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge. https://doi.org/10.4324/9780203841792

Skehan, P. (2009). Lexical performance by native and non-native speakers on language-learning tasks. In B. Richards, H. Daller, D. D. Malvern, & P. Meara (Eds.), *Vocabulary studies in first and second language acquisition: The interface between theory and application* (pp. 107–124). Palgrave Macmillan. https://doi.org/10.1057/9780230242258_7

Skehan, P. (2014a). Introduction: Coordinated research into task-based performance. In P. Skehan (Ed.), *Processing perspectives on task performance*. (pp. 1–26). https://doi.org/10.1075/tblt.5.01ske

Skehan P. (2018). *Second language task-based performance: Theory, research, and assessment*. Routledge. https://doi.org/10.4324/9781315629766

Skehan, P. (2022a). The psycholinguistics of task-based performance. In M. Long & M. Ahmadian (Eds.), *The Cambridge handbook of task-based learning and teaching* (pp. 3–26). Cambridge University Press. http://doi.org/10.1017/9781108868327.002

Skehan, P. (2022b). Performance on second language speaking tasks: Supports and impediments. In Schwieter, J. & Benati, A. (Eds.), *Second Language Acquisition Theory: The legacy of Professor Mike Long* (pp. 211–234). John Benjamins. https://doi.org/10.1075/bpa.14.11ske

Skehan, P., Bui, G., Wang, Z. & Shum, S. (2024). Re-examining accuracy measures in task-based second language performance. *Research Methods in Applied Linguistics*, *3*(1), 100098. https://doi.org/10.1016/j.rmal.2024.10098

Skehan, P., & Foster, P. (1997). The influence of planning and post-task activities on accuracy and complexity in task-based learning. *Language Teaching Research*, *1*(3), 185–211. https://doi.org/10.1177/136216889700100302

Skehan P. & Foster P. (1999). Task structure and processing conditions in narrative retellings. *Language Learning*, *49*(1), 93–120. https://doi.org/10.1111/1467-9922.00071

Skehan, P. & Foster, P. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 193–216). John Benjamins B.V. https://doi.org/10.1075/lllt.11.12ske

Skehan, P., & Shum, S. (2014). Structure and processing condition in video-based narrative retelling. In P. Skehan (Ed.), *Processing perspectives on task performance*. (pp. 197–210). https://doi.org/10.1075/tblt.5.07ske

Tavakoli, P., & Wright, C. (2020). *Second language speech fluency: From research to practice*. Cambridge University Press. https://doi.org/10.1017/9781108684467

Wang, Z. (2009). *Modelling speech production and performance: Evidence from five types of planning and two task structures.* Unpublished Ph.D. dissertation, Chinese University of Hong Kong.

Wang, Z. (2014). On-line time pressure manipulations: L2 speaking performance under five types of planning and repetition conditions. In P. Skehan (Ed.), *Processing perspectives on task performance*. (pp. 27–62). https://doi.org/10.1075/tblt.5.02wan

Wang, Z., & Skehan, P. (2014). Task structure, time perspective and lexical demands during video-based narrative retellings. In P. Skehan (Ed.), (pp. 155–186). https://doi.org/10.1075/tblt.5.06wan

Appendix One

Worked examples of length-linked accuracy measures

Calculations for the accuracy measures that are linked to length-of-clause are described below. Calculations do not include one-word utterances. Table 5 presents the data on which the calculations are based. The first column shows the number of words in each clause, ranging from 2 to 10 (this being the maximum in the example data). The second and third columns present the basic data, with the number of correct clauses in column 2 (for each clause length) and the number of incorrect clauses in column 3. For example, there are six correct two-word clauses and three incorrect two-word clauses.

Error-free Clauses (EFC) and Error-free Clauses > 4 (EFC>4): The calculations of these scores are based on the data in Columns 2 and 3. For EFC, the correct clause total (39) is divided by the total clauses, the sum of correct and incorrect clauses (39+64), giving the value 0.61. In slight contrast, for EFC>4, the data from the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

Table 5. Example data and calculations for EFC, EFC>4, Prop>4, and WeightedNetAccuracy

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8
No. of words	EFC EFC>4		Prop>4		Acc.Lambda	Weighted Net-Accuracy	
	Correct	Incorrect	Proportion	Multiple	Calculation	Base calculation	Multiple
2	6	3	(0.67)		2 × 6 = 12	2 × (6−3)	6
3	8	2	(0.80)		3 × 8 = 24	3 × (8−2)	18
4	7	2	(0.77)		4 × 7 = 28	4 × (7−2)	20
5	6	3	0.66	3.30	5 × 6 = 30	5 × (6−3)	15
6	5	3	0.63	3.78	6 × 5 = 30	6 × (5−3)	12
7	4	4	0.50	3.50	7 × 4 = 28	7 × (4−4)	0
8	2	5	0.28	2.24	8 × 5 = 40	8 × (2−5)	−24
9	1	3	0.25	2.25	9 × 1 = 9	9 × (1−3)	−18
10	0	0	–		–	–	
Sums	EFC			15.07	201		29
	39	25					
	EFC>4						
	18	36					
Score	EFC 39/64 = 0.61 EFC>4 18/36 = 0.50		15.07/5 = 3.01		201/8=25.13	29/8 = 3.63	

first three rows (clauses of length 2, 3, 4) are omitted, and so the relevant totals are 18 and 36, thus giving the value 0.50.

Proportion > 4 (Prop>4): First of all, the data from clause lengths 2, 3, and 4 words is omitted. Then column 4 shows the proportion correct for the clause lengths 5 to 10, such as, for 5-word clauses, 6/3 (from columns 2 and 3), giving the value 0.66. This is then multiplied by the clause length, 5, to give the value 3.30. The same is done for clause lengths 6 to 10, generating the values 3.78, 3.50, 2.24, and 2.25. The five product terms are then summed, giving the value 15.07, and this total is divided by the number of terms, 5, to give the average value of 3.01.

AccuracyLambda: Calculations for this measure are shown in column 6. These calculations are based on all clause lengths, but only use the numbers of correct clauses. Each clause length in words is multiplied by the number of correct clauses, as with 2 x 6 for the two-word clauses. These product terms are then summed, in the example giving 201, and then this is divided by the number of product terms used, in this case giving 201/8, which equals 25.13.

The WeightedNetAccuracy score has similarities in computation, but the key difference is that it is the net score that is calculated for each clause length, i.e. correct minus incorrect clauses. The first term would therefore become 6-3, giving the result 3. Then, as with Accuracy Lambda, terms are summed and divided by the total number of product terms. In the example case, the relevant total is 29, which is divided by 8, giving the result 3.63. As noted in the text earlier, with each clause length, if the number of incorrect clauses exceeds the number of correct clauses, the product term becomes negative, and this, clearly, has an impact on the total that is calculated.

For the Length Accuracy 70% score, all clauses are ranked in length. Then the proportion correct for each clause length is calculated, as shown in column 4 in Table 5, and now we include the 2, 3, and 4-word clauses, where the proportions are bracketed. These values are examined, and the greatest length that can be handled with 70% accuracy or better is assigned as the achieved score. If we take the above data, for reasons of ease of calculation, we will make one change. From the data we have been using so far, we will change the scores for two-word clauses from 6 correct and three incorrect to 7 correct and 3 incorrect (the reason for this will become clear below). This would give the proportions 2 words: 70%, 3: 80%; 4: 78%; 5: 66%, 6: 63%, 7: 50%; 8: 40%; 9: 25%. This would give a neat cut-off score at the four-word clause level, in that clause lengths of 2-4 meet the 70% criterion, but those above it do not. But of course, in this example, there has been cheating! If we return to the original scores, we actually have 2: 66%, 3: 80%; 4: 78%; 5: 66%, 6: 63%, 7: 50%; 8: 40%; 9: 25%. So,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892

two-word clauses actually fail the 70% criterion, but three- and four-word clauses meet the standard. One can ask therefore whether the 2-word result was a blip in performance and unrepresentative, and the longer clauses more validly capture this candidate's actual level. In the main text, we discuss the way problems such as this (and there are more complex patterns of scores than this) might require more complex supplementary scoring rules, but might introduce some unreliability to the judgments which are made with this particular index. The rule devised for the present example is that two consecutive successful performances overcome a blip at just one level (but not two such blips).

Cite this article: Skehan, P., & Wang, Z. (2026). Exploring new measures of accuracy in second language task-based spoken performance. *Studies in Second Language Acquisition*, 1–20. https://doi.org/10.1017/S0272263126101892

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 12:49:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0272263126101892