

Task Performance Measures: A Cognitive Perspective

Qian Li¹

¹ Center for Lexicographical Studies, Guangdong University of Foreign Studies, Guangzhou, China

Correspondence: Qian Li, Center for Lexicographical Studies, Guangdong University of Foreign Studies, Guangzhou, China.

Received: November 2, 2020

Accepted: December 16, 2020

Online Published: December 21, 2020

doi:10.20849/jed.v4i3.837

URL: <https://doi.org/10.20849/jed.v4i3.837>

Abstract

To explore more about task performance in second language acquisition research, how task performance is measured turns out to be an indispensable topic with much concern from different perspectives. The present study takes the cognitive perspective to analyze the measures of the different aspects of task performance, i.e. fluency, complexity, accuracy and lexical performance. It is revealed that for the same construct of performance measures, different features are involved in different indexes. Multivariate measures of the same construct need to be adopted to seek an overall picture of learners' task performance.

Keywords: fluency, complexity, accuracy, lexical performance

1. Introduction

In task-based research, how task performance is measured stands as a topic with crucial significance. In the literature, task performance measures vary greatly among task researchers. To a considerable extent, the different measures of task performance adopted in different studies have reflected researchers' theoretical positions (Skehan, 2003). Interactionists focus on the indices for the negotiation of meaning, such as clarification requests, confirmation checks, comprehension checks, recasts and uptakes (Russell & Spada, 2006); Researchers based on the sociocultural theory prefer to the measures of interactive and language involvement, such as language-related episodes (LREs) and turns (Lowen, 2005). Working in the framework of the cognitive approach, researchers distinguish oral performance in the following directions: fluency, accuracy, complexity, and recently the lexical aspects of performance (Skehan, 1998; Skehan & Foster, 2005; Yuan & Ellis, 2003).

The present study focuses on the cognitive perspective for task performance measures. Specifically, the measures of complexity, accuracy, fluency and lexical performance are taken under exploration in this study. However, even for the same construct of task performance (e.g. complexity), different measures are used across studies in the literature (Ortega, 2003).

2. Conceptual Background

In the traditional sense, some researchers argue that the three components of language performance (i.e. fluency, complexity and accuracy) reflect two different aspects of language processing: language representation (also known as declarative knowledge) and language access (also known as procedural knowledge) (Wolfe-Quintero et al, 1998). In this view, complexity and accuracy correspond to the language representation of L2 learners. Particularly, complexity reveals "the scope of expanding or restructured knowledge to target language norms", accuracy shows "the conformity of L2 knowledge to target language norms" (Wolfe-Quintero et al, 1998, p 4). In comparison, fluency reflects how learners control the access of the language, "with control improving as the learner automatizes the process of gaining access" (ibid, p.4).

From an information processing perspective, Skehan (1998) argues that the three aspects—fluency, complexity and accuracy—all pertain to L2 language representation, but language representation consists of dual systems, and also different aspects of language performance draw on different linguistic systems: the exemplar-based and the rule-based systems. To achieve fluency, learners make use of the exemplar-based (or memory-based) system which consists of formulaic chunks for an increased production speed. In contrast, accuracy and complexity require learners to draw on the rule-based system in which the abstract rules can be used to create a variety of utterances/sentences. According to Skehan(1998), in task-based language teaching and learning, fluency refers to real-time language production without undue pausing or hesitation which occurs when learners take the meaning as the primary concern to get the task done; complexity reflects learners' willingness to try the interlanguage

structures that are ‘cutting edge’ and elaborated; and accuracy refers to how well the target language is produced in relation to the rule system of the target language. Based on the different primacies of the three aspects, Skehan proposed that in language production, there would be an initial contrast between meaning (fluency) and form (complexity and accuracy). Further, the nature of complexity (i.e. restructuring of language) and of accuracy (i.e. control of language) can represent another contrast between them, although both complexity and accuracy are relevant for the rule-based system and viewed as the formal aspects of performance (Skehan, 1998).

Independent of the above syntactic aspects, lexical performance has attracted attention in a relatively small number of task-based studies (Skehan, 2009). Although vocabulary is treated as an indispensable aspect of language knowledge (Daller, Milton, & Treffers-Daller, 2007), lexical performance has not received sufficient attention in the task-based research until recently. In their research concerning the measures of L2 development, Wolfe-Quintero et al (1998) include the lexical aspects as independent of fluency, accuracy and grammatical complexity and argue that lexical richness constitutes another major area for language complexity. According to Wolfe-Quintero et al, lexical complexity is manifested in terms of the range (lexical variation) and size (lexical sophistication) of L2 vocabulary. Daller et al (2007) proposed the multiple measures of different aspects of lexical performance and viewed lexical richness as an umbrella term covering lexical diversity, which is “the variety of active vocabulary deployed by a speakers or writer” (Malvern & Richards, 2002: 87), lexical sophistication (the number of low frequency words) and lexical density (the ratio of content and function words) (Daller et al, 2007:13).

3. Measures of Fluency, Complexity, Accuracy and Lexical Performance

3.1 Measures of Fluency

Speech fluency is of such a multifaceted nature that it manifests and reflects not only underlying speech-planning and thinking processes but also the process of speech production, the phenomenon of hesitation and the temporal dimensions of speech (Freed, 2000). Lennon (1990) viewed oral fluency measures as two types: measures of temporal aspects, such as words per minute or pause length, and measures of dysfluencies, such as repairs. In the literature of SLA research, a comprehensive cluster of fluency-related measures were used including amount of speech, rate of speech, unfilled pauses, frequency of filled pauses, length of fluent speech runs, repairs (including repetitions, reformulations/false starts, corrections and partial repeats), and clusters of dysfluencies (Freed, 2000). However, research shows that not all the fluency measures are necessarily effective in distinguishing the different oral proficiency levels, that is, not all the measures are perfectly reliable in oral data analysis (Freed, 2000). Pauses, for instance, may reflect either time required to focus on a new thought or time required to put a thought into words (Lennon, 1990). Thus, “the presence of pauses is not exclusively associated with a lack of fluency in a second language” (Freed, 2000:256). In the present study, fluency, only some of potential fluency measures are adopted as follows. Given the multi-functionality of pauses in L2 speech, pauses are not included in the fluency measure. The adopted fluency measures are:

- a. Speech rate: refers to how fast and dense the produced language is in terms of the time units. It is calculated on the basis of the number of (nonrepeated) words per minute;
- b. Repair fluency: including the following categories: reformulation which suggests a decision to rephrase the form syntactically or morphologically, replacement which reflects the change of vocabulary in speech, false starts which occur where an utterance is begun and then is abandoned, and repetition which means the repetition of a word or a string of words (Foster, Tonkyn & Wigglesworth, 2000; Skehan, coding manual).
- c. Filled pauses: refers to the non-lexical fillers in the speech.

The above three kinds of fluency measures reflect both the temporal measures and the dysfluency aspects of oral performance which may reveal an overall picture of participants’ fluency.

3.2 Measures of Complexity

The following performance aspects are all form-related. The measure of complexity is closely related with the various ways of segmenting language performance into units. Segmenting the language performance based on T-units is widely used in the analysis of written language. Hunt (1965) first developed the T-unit to measure children’s syntactic maturity in writing, and defined the T-unit as a “minimal terminable unit” consisting of a main (i.e. independent clause) plus any subordinate clauses. Later, the definition of T-unit was adopted in the studies of spoken language as well. However, the T-unit has been criticized for the following reasons: first, the definition of T-unit only includes subordination, but not coordination, thus inappropriately dividing coordinations into different T-units (Bardovi-Harlig, 1992); second, the T-unit, originally developed for the analysis of written language, is not suitable for the analysis of spoken data which consist often of many

fragments and elliptical sentences (Foster et al, 2000). Alternatively, for spoken language analysis, some researchers adopted the communication unit (C-unit). The C-unit refers to not only grammatical independent predications but answers to questions which lack only the repetition of the question elements. For example, according to Loban, “Yes” can be viewed as a whole unit of communication when it is an answer to a question such as “have you ever been sick?” However, Foster et al (2000) pointed out that the definition of C-unit seems to exclude “elliptical constructions which arise within a speaker’s turn rather than link to an interlocutor’s question” (p. 361) such as the topical noun phrase. It seems, therefore, that neither T-unit nor C-unit is suitable for the analysis of spoken data.

Foster et al (2000) proposed the “analysis of speech unit” (AS-unit) which can be applied reliably to spoken language. “An AS-unit is a single speaker’s utterance consisting of an independent clause or sub-clausal unit, together with any subordinate clause(s) associated with either” (Foster et al, 2000: 365). An independent clause will be minimally a clause including a finite verb. As an analysis unit for spoken data, AS-unit also includes independent sub-clausal units which consist of either one or more phrases which can be elaborated to a full clause or a minor utterance which will be defined as irregular sentences or nonsentences, such as “thank you” “oh poor woman”. (Foster et al, 2000:365-366). In this way, in contrast to the T-unit, fragments which are common in speech are included in AS-units. In addition, the AS-unit takes other features of spoken data into consideration, such as dysfluency features, topicalization, interruption and scaffolding.

There are various types of grammatical complexity ratios. One type is the general complexity measure (clauses per production unit) which considers the proportion of all clause types to a larger unit. Another type is the dependent clause measure (dependent clauses per clause or per production unit etc) which considers the relationship between dependent and independent clauses. A third type is the coordination measure (coordinate clauses per clause or production unit etc) which considers the relationship between the coordination and independent clauses (Wolfe-Quintero et al, 1998). As compared to the dependent and coordination measures, a general complexity measure has been used in a larger number of studies, albeit with mixed findings (Wolfe-Quintero et al, 1998). The present research follows those previous studies to use a general complexity measure. In particular, complexity is measured by dividing the total number of clauses by the total number of AS-units. One crucial issue in the measure of grammatical complexity in spoken data is how to deal with fragments by researchers. As Foster et al (2000) proposed, fragments are included in AS-units.

In addition to the AS-unit complexity ratio, the mean length of AS-unit is considered as well. Although some researchers viewed the mean length of production unit as a fluency measure (Wolfe-Quintero et al, 1998), it is accepted among most researchers that the mean length of production unit is a measure of complexity (Norris & Ortega, 2008) and the measure has been sufficiently investigated across studies.

As Norris & Ortega (2008) proposed that since syntactic complexity is multi-faceted, measures of complexity need to be multivariate. In past studies, the length of production units (e.g. T-units) and the number of clauses per T-unit are found to be the best ways to predict learner proficiency and also had a significant linear relation with independent oral proficiency measures (Iwashita, 2006). To adopt more varied and valid measures of complexity for spoken data, the present research, therefore, uses the two effective measures of syntactic complexity—the mean length of AS-unit and the number of clauses per AS-unit.

3.3 Measures of Accuracy

Studies of second language acquisition have used various measures of accuracy. Some researchers measure accuracy by looking at the target-like usage of specific language features, such as past tense morphemes, plural morphemes, the number of correct pronouns or correct definite articles (Ortega, 1999). However, it has also been argued that such specific measures are less sensitive to detecting differences between experimental conditions.

The majority of researchers measure language accuracy by taking general errors into account. Two approaches are used in the literature. One is to focus on whether a structural unit (e.g. clauses, sentences, T-units or AS-unit) is error-free. Typical measures are the number of error-free T-units per T-unit, or the number of error-free clauses per total clauses, or the number of error-free clauses per T-unit (Wolfe-Quintero et al, 1998). Error-free measures have also been criticized for the following problems: first, they are concerned with the quantities of error-free strings, but not the quality, in other words, this error-free measure does not reveal what types of errors are involved (Wolfe-Quintero et al, 1998); second, this measure does not reveal how errors are distributed within a unit, because a unit containing a single error is treated equally as a unit containing multiple errors; third, this error-free measure does not take the length of the analysis unit into account. Then, a high error-free ratio would be misleading when learners produce a large number of short but accurate analysis units (Skehan & Foster, 2005).

In view of the above criticism, researchers have developed other ways concerning accuracy measures. Some have proposed the calculation of errors in relation to production units (such as the number of errors per word, or per T-unit etc) (Kepner, 1991). This approach is concerned with the quantities of errors, and it is better in distinguishing production units with one error or more than one error than the error-free measure does. Further, Bardovi-Harlig & Bofman (1989) propose the use of clauses rather than T-units in the measure of accuracy in order to eliminate complexity as a factor. Other researchers improved the error-free measure to be more sensitive to the length of error-free clauses (Skehan & Foster, 2005). In addition to the ratio of error-free clauses, Skehan & Foster suggest calculating the error-free proportion in different clause lengths, that is to calculate the error-free proportion among three-word clauses, then among four-word clauses, and so on. Ideally, this would reveal a cut-off point, beyond which the participant cannot produce the correct clauses to meet a required criterion level. As for the criterion level, they suggest the 50%, 60% and 70% criterion levels be the likely candidates used in the task-based research (Skehan & Foster, 2005).

Bearing in mind the defects of the different accuracy measures and the corresponding developments, the current study takes the following accuracy measures collectively: first of all, the error-free clause ratio is calculated by dividing the number of error-free clauses by the total number of clauses without the interference of AS-unit segmentation; in addition, the error-free clause ratio is further calculated in different clause lengths, and a 70% criterion is adopted in the study to discriminate various accuracy levels of participants; further, to eliminate the defects of error-free measures, errors per 100 words is counted. Errors are defined as any deviation from the standard in terms of morphological, syntactic and lexical aspects.

3.4 Measures of Lexical Performance

With regard to lexical performance, many commonly used measures have been based on the ratio of different words (types) to the total number of words (tokens), known as the Type-Token Ratio (TTR). The TTR has been criticized because it is sensitive to text length (Wolfe-Quintero et al, 1998). There is a negative relationship between a type/token ratio and sample size, that is samples with larger numbers of tokens give lower values for TTR and vice versa, because the longer samples of language that are produced, the more of the active vocabulary is likely to be included and the available pool of new word types that can be introduced steadily diminishes (Malvern & Richards, 2002). Given that most task performances of second language learners are fairly short, the TTR for L2 learner's lexical performance poses acute problems (Skehan, 2003). According to Read (2000), the TTR should be considered as a single measure of lexical diversity. Nation (2007) strongly argues that since vocabulary knowledge is multi-dimensional, it is necessary to adopt a set of complementary measures that reveal different aspects of vocabulary knowledge. In attempt to arrive at such multiple measures, researchers proposed the concept of "lexical richness" which includes different aspects of vocabulary use, such as lexical variation (i.e. lexical diversity), lexical density and lexical sophistication (Daller et al, 2007). In their synthesis research, Wolfe-Quintero et al (1998) find that lexical variation and lexical sophistication are related to language development, but not lexical density. In the field of vocabulary assessment and measures, researchers pay more attention to lexical sophistication and lexical variation than to lexical density (see Daller et al, 2007). Following the past studies, the present research takes lexical variation and lexical sophistication as the methods of measuring lexical performance.

Lexical variation (also lexical diversity) refers to the variety of learners' vocabulary. As an alternative to the type/token ratio, *vocd* (Note 1) was developed by Malvern & Richards (2002) which assumes another functional relationship between the number of types and tokens, thus providing a new measure of vocabulary diversity referred as the *D* value. Put simply, *D* provides an index of the extent to which the speaker avoids the recycling of the same set of words. A lower *D* suggests a larger tendency to return to a set of words by the speaker. Thus, *D* is a text-internal measure of lexical performance (Skehan, 2009). As compared with previous measures, *D* has been proven to be better in that it may avoid the flaw in raw TTR (Malvern & Richards, 2000). However, *D* is not perfect in that as a type/token-based measure it does not take into account the frequency of a word. A common word and a rare one have the same weights in the measure of lexical variation.

Unlike *D* and other type-token based measures, the measure for lexical sophistication is based on the notion of frequency. Lexical sophistication is measured by determining the number of low frequency words in a text (Read, 2000). *P-Lex*, for example, as a computer program can do this automatically by dividing a text into ten word chunks and then calculating the number of infrequent words in each ten word chunk (Skehan, 2009). *P-Lex* is based on a probability distribution (Poisson distribution) that is taken as a model for the occurrence of rare or difficult words. To define the rare or difficult words, a word frequency list is needed in *P-Lex*. As Daller & Xue (2007) pointed out, the word lists for *P-Lex* "have to be chosen carefully and have to be adapted to the specific task" (p.164). For the narrative and the decision-making tasks used in the present study, *P-Lex* was rewritten by

Skehan with the reference to the British National Corpus spoken component. In this modified program, the word list is lemmatised. Files of task-specific words are compiled to enable words to be temporarily defined as easy. In addition, a cut-off value of fewer uses than 150 per million words is used to define words difficulty (Skehan, 2009). “This value seemed to be most effective in producing a good range of discrimination” (Skehan, 2009: p.110).

The two kinds of measures for lexical variation and lexical sophistication are complementary to each other—the frequency-based measure of lexical sophistication is a text-external measure and reveals “the access to and deployment of more rare or more difficult, more precise vocabulary” (Richards & Malvern, 2007:84), while the type-token based measure of lexical variation reflects the “access to a wide range of vocabulary, and by inference, its skillful use” (Richards & Malvern, 2007:84). The present study adopts the two measures of lexical richness in hope to give a complete picture of lexical performance in different experimental groups.

4. Conclusion

From the cognitive perspective, task performance is generally measured in terms of fluency, complexity, accuracy, and lexical performance (i.e. lexical variation and lexical sophistication). To tap into the different features of a performance construct, multivariate measures are adopted for the same construct. Those measures may be independent of each other, but they are complementary to each other to provide an overall picture of learners’ task performance.

References

- Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly*, 26, 390-395. <https://doi.org/10.2307/3587016>
- Bardovi-Harlig, K., & Bofiman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition*, 11, 17-34. <https://doi.org/10.1017/S0272263100007816>
- Daller, H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *A causal experimental study of the reciprocal relationship between vocabulary knowledge, IELTS, and academic grades* (pp. 150-164). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511667268.011>
- Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Fundamental issues in modelling and assessing vocabulary knowledge* (pp. 35-43). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511667268>
- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21, 354-375. <https://doi.org/10.1093/applin/21.3.354>
- Freed, B. F. (2000). Is Fluency, Like Beauty, in the Eyes (and Ears) of the Beholder?. In H. Riggensbach (Ed.), *Perspectives on fluency* (pp. 243-265). Anna Arbor: University of Michigan Press.
- Hunt, K. W. (1965). *Grammatical structures written at three grade levels*. Urbana, IL: The National Council of Teachers of English.
- Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Language Assessment Quarterly*, 3, 151-169. https://doi.org/10.1207/s15434311laq0302_4
- Kepner, C. G. (1991). An experiment in the relationship of types of written feedback to the development of second-language writing skills. *Modern Language Journal*, 75, 305-313. <https://doi.org/10.1111/j.1540-4781.1991.tb05359.x>
- Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40, 387-417. <https://doi.org/10.1111/j.1467-1770.1990.tb00669.x>
- Loewen, S. (2005). Incidental focus on form and second language learning. *Studies in Second Language Acquisition*, 27, 361-386. <https://doi.org/10.1017/S0272263105050163>
- Malvern, D., & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan, & A. Wray (Eds.), *Evolving models of language: Papers from the annual meeting of the British Association for Applied Linguistics* (pp. 58-71, September, 1996). Held at the University of Wales, Swansea,. Clevedon, UK: Multilingual Matters. <https://doi.org/10.1191/0265532202lt221oa>
- Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19, 85-104.

- Norris, J., & Ortega, L. (2008). *Measurement for understanding: The case of complexity, accuracy and fluency*. AAAL 2008, Washington DC.
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24, 492-518. <https://doi.org/10.1093/applin/24.4.492>
- Read, J. (2000). *Assessing vocabulary*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511732942>
- Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar: A meta-analysis of the research. In J. M. Norris, & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp.133-162). John Benjamins Publishing Company. <https://doi.org/10.1075/llt.13.09val>
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press. <https://doi.org/10.1177/003368829802900209>
- Skehan, P. (2003). Task-based instruction. *Language Teaching*, 36, 1-14. <https://doi.org/10.1017/S026144480200188X>
- 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, J. Milton, B. Richards, & J. T. Daller (Eds.), *Vocabulary studies in first and second language acquisition: The interface between theory and application* (pp.107-124). London: Palgrave Macmillan. https://doi.org/10.1057/9780230242258_7
- Skehan, P., & Foster, P. (2005). *Strategic and on-line planning: The influence of surprise information and task time on second language performance*. Amsterdam: John Benjamins. <https://doi.org/10.1075/llt.11.12ske>
- Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.
- Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 oral performance. *Applied Linguistics*, 24, 1-27. <https://doi.org/10.1093/applin/24.1.1>

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).