





METHODS FORUM

Evaluating evidence for the reliability and validity of lexical diversity indices in L2 oral task responses

Kristopher Kyle¹ , Hakyung Sung¹, Masaki Eguchi¹  and Fred Zenker² 

¹Department of Linguistics, University of Oregon, Eugene, OR, USA; ²Department of Second Language Studies, University of Hawaii at Manoa, Honolulu, HI, USA

Corresponding author: Kristopher Kyle; Email: kkyle2@uoregon.edu

(Received 04 July 2022; Revised 16 June 2023; Accepted 12 July 2023)

Abstract

Although lexical diversity is often used as a measure of productive proficiency (e.g., as an aspect of lexical complexity) in SLA studies involving oral tasks, relatively little research has been conducted to support the reliability and/or validity of these indices in spoken contexts. Furthermore, SLA researchers commonly use indices of lexical diversity such as Root TTR (Guiraud's index) and *D* (vocd-*D* and HD-*D*) that have been preliminarily shown to lack reliability in spoken L2 contexts and/or have been consistently shown to lack reliability in written L2 contexts. In this study, we empirically evaluate lexical diversity indices with respect to two aspects of reliability (text-length independence and across-task stability) and one aspect of validity (relationship with proficiency scores). The results indicated that neither Root TTR nor *D* is reliable across different text lengths. However, support for the reliability and validity of optimized versions of MATTR and MTLD was found.

Indices of lexical diversity (and in particular indices of lexical variety) are often used as measures of lexical proficiency and/or lexical development in studies of second language acquisition (SLA; Bulté & Roothoof, 2020; Lambélet, 2021; Tracy-Ventura et al., 2021; Vidal & Jarvis, 2020). As language learners develop (and become more proficient language users), we expect that the size of their productive vocabulary will grow. Accordingly, given a particular language production task, we would expect that more proficient language users would use a wider variety of lexical items to complete the task. We also presume that more proficient language users would produce longer texts than less proficient users (Carlson et al., 1985; Iwashita et al., 2008; Jarvis et al., 2003). A well-known, but often ignored, issue with indices of lexical variety is that many have been shown to be strongly (and intrinsically) related to text length (Hess et al., 1986; McCarthy & Jarvis, 2010; Zenker & Kyle, 2021). Some indices, such as the type-token ratio (TTR), are intrinsically *negatively* correlated with text length. This is problematic because TTR scores decrease as texts become longer—in other words,

more fluent speakers and writers earn lower lexical diversity scores. Other well-known indices, such as Root TTR (Guiraud's index), are intrinsically *positively* correlated with text length. As text lengths increase, Root TTR also increases. This is undesirable because it means that Root TTR conflates text length and lexical diversity: when a positive relationship between proficiency or development and Root TTR is found, it is unclear whether the observed relationship is due to increases in lexical diversity or productivity (e.g., the total number of words produced; Norris & Ortega, 2009).

In acknowledgement of the intrinsic relationship between TTR and text length, many studies have attempted to develop text-length independent measures of lexical diversity. Although many early attempts have been shown to be problematic (Chotlos, 1944; Guiraud, 1960; Maas, 1971; Malvern & Richards, 1997), more recent proposals have shown promise (Covington & McFall, 2010; McCarthy & Jarvis, 2010). For SLA researchers, one potential issue with extant studies is that they have tended to focus on longer L1 texts (McCarthy & Jarvis, 2007, 2010). A few studies have used shorter written L2 texts (Vidal & Jarvis, 2020; Zenker & Kyle, 2021), but in the realm of L2 speech only small-scale studies have been conducted (Koizumi & In'nami, 2012). In this study, we extend previous L2 studies by examining the degree to which indices of lexical variety are stable across varying text lengths and across task types in a large corpus of L2 oral proficiency interviews.

Lexical diversity indices and text-length stability

As with any other construct we want to measure, an index of lexical diversity should be both demonstrably reliable and arguably valid. Because learners may create productions of different lengths, even when timed tasks are used, indices of lexical diversity need to be reliable (i.e., consistent) across texts of different lengths. Accordingly, there has been a particular focus on text-length stability in the literature (Guiraud, 1960; Hess et al., 1986; Jarvis, 2002; Koizumi & In'nami, 2012; McCarthy & Jarvis, 2010; Tweedie & Baayen, 1998; Vidal & Jarvis, 2020; Zenker & Kyle, 2021).

It should be noted that it is not necessarily problematic for indices of lexical diversity to be correlated with text length in a particular corpus. We expect a positive correlation between proficiency (broadly construed) and both lexical diversity and fluency (i.e., temporal aspects of speech; Lennon, 2000) as well as productivity (the total amount of production; Norris & Ortega, 2009), and therefore correlations between lexical diversity and these constructs are expected. It is, however, problematic when lexical diversity indices *intrinsically* vary due to text length. One common method of determining the degree to which indices vary intrinsically due to text length is the parallel sampling method (Hess et al., 1986), which involves dividing a text into sections of a particular length and then averaging index scores across the sections. When the parallel sampling method is repeated with sections of several different lengths, correlations can be calculated between section length and lexical diversity scores. Zenker and Kyle (2021), for example, analyzed the relationship between text length and lexical diversity indices in a large corpus ($n = 4,542$) of L2 argumentative essays (i.e., the ICNALE corpus; Ishikawa, 2011). They divided each text into subsections from 50 to 200 words in length in five-word increments (50 words, 55 words, 60 words, etc.). Among the nine indices of lexical diversity examined, moving average TTR (MATTR) was the most reliable across different text lengths while TTR, Root TTR, and Log TTR were the least reliable (and were strongly related to text length). Similar, relatively large-scale analyses have been conducted in other L2 writing contexts (Vidal & Jarvis, 2020) and using L1 corpora consisting of longer spoken and written texts (McCarthy & Jarvis, 2007, 2010).

To our knowledge, only small-scale studies have investigated text-length reliability in the types of spoken L2 contexts that are common in SLA research. Koizumi and In'nami (2012), for example, analyzed monologic speaking task responses from 38 participants. Using the parallel sampling method on text samples of 50–200 words in length, they found that none of the lexical diversity indices examined (including Root TTR and *D*) were completely independent of text length, though MTLT stabilized at 100 words. Although their results are helpful, more robust analyses are needed to make strong claims related to the reliability of measures.

Of the indices of lexical diversity that have been proposed, at least five merit some discussion due to their conceptual influence on the field of SLA (TTR), their continued use by SLA researchers (Root TTR), and/or their potential promise (*D*, MATTR, MTLT). Perhaps the most well-known index of lexical diversity is the type-token ratio (TTR), which is calculated by dividing the number of unique words in a text (i.e., the number of types) by the number of total running words in a text (i.e., the number of tokens). Although this index is conceptually straightforward, it has a well-known and critical weakness—namely, that it varies intrinsically due to text length (Guiraud, 1960; McCarthy & Jarvis, 2010; Tweedie & Baayen, 1998; Zenker & Kyle, 2021). As texts get longer, both function words and content words tend to be repeated until a topic shift occurs, at which point new content words might be introduced (though function words still tend to be repeated). Consequently, when TTR is used as an index of lexical diversity, it tends to overestimate the diversity of shorter texts (typically by less proficient users) and underestimate the diversity of longer texts (typically by more proficient users). An early attempt to mitigate this relationship was to transform the TTR value by using the square root of the number of tokens in the denominator (Guiraud, 1960). Although this index, commonly referred to as Root TTR or Guiraud's index, gained a reputation as an appropriate substitute for TTR and is still used fairly widely (Bulté & Housen, 2019; Lambelet, 2021), studies have repeatedly shown that it strongly overcorrects TTR's negative relationship with text length (e.g., Koizumi & In'nami, 2012; McCarthy & Jarvis, 2010; Zenker & Kyle, 2021). One possible reason for the durability of Root TTR in the field is that it tends to demonstrate a relatively strong relationship with proficiency, and it is certainly an improvement over TTR because it is positively correlated with text length and therefore does not penalize longer essays (see Bulté & Roothoof, 2020; Treffers-Daller *et al.*, 2018). However, because the index is intrinsically positively correlated with text length, it is unclear to what degree increases in Root TTR scores can be attributed to increases in lexical diversity, fluency, productivity, and/or topic development, among other potential causes for increases in text length.

Although previous research has clearly demonstrated that TTR and Root TTR are intrinsically related to text length, at least three other indices of lexical diversity have been proposed in recent decades that show more promise regarding text-length independence. The first is the index *D*, which is most commonly operationalized as *vocd-D* (Malvern *et al.*, 2004; Malvern & Richards, 1997) using software such as CLAN (MacWhinney, 2000). The index *vocd-D* attempts to measure lexical diversity while mitigating text-length effects using a bootstrapped approach that repeatedly fits the rate of decline in TTR values within texts (i.e., the TTR curve) in random samples of varying lengths from a text. McCarthy and Jarvis (2007) demonstrated that *D* can be calculated in a more precise and straightforward manner by simply calculating the probability that each word type in a text would occur in a random sample from the text (using a hypergeometric distribution). McCarthy and Jarvis (2010) and subsequent studies have referred to this index as HD-*D*. Results with respect to text-length stability for *D* have been mixed, though most studies have found at least a small relationship between *D* and text length (Malvern & Richards,

1997; McCarthy & Jarvis, 2007; Zenker & Kyle, 2021). There is evidence, however, that *D* may be more sensitive to text length in spoken texts (McCarthy & Jarvis, 2007) and that these effects may be large (Koizumi & In'nami, 2012).

The second index that has shown promise is the Measure of Textual Lexical Diversity (MTLD; McCarthy & Jarvis, 2010). The MTLD leverages the relationship between TTR values and text length by calculating how quickly TTR values stabilize. MTLD scores represent the average number of words it takes for TTR values to fall to the point of stabilization in a text (usually operationalized as $TTR = .720$). Texts with larger MTLD values are considered more diverse because it takes longer before enough word repetition occurs for TTR values to drop to the stabilization point. Research with L1 spoken and written registers (McCarthy & Jarvis, 2010) and L2 argumentative essays (Vidal & Jarvis, 2020; Zenker & Kyle, 2021) has suggested that MTLD values tend to be resistant to text length effects (but see Treffers-Daller, 2013). There is also preliminary evidence that MTLD is reasonably stable across some L2 spoken registers (Koizumi & In'nami, 2012), though more robust sample sizes are needed to confirm this. One well-documented weakness of MTLD is the estimation of partial factors (text segments at the end of a production that are longer than 10 words but have not yet reached the cut-off TTR value). The default solution has been to average scores (including partial factors) for MTLD calculated forward and backward through the text, though windowed approaches have also been used (see Vidal & Jarvis, 2020; Zenker & Kyle, 2021). Use of MTLD is increasingly common in SLA studies (e.g., Bulté & Roothoof, 2020; Pfenniger, 2020; Vidal & Jarvis, 2020).

The third index of promise is moving-average TTR (MATTR; Covington & McFall, 2010). MATTR mitigates the relationship between TTR and length by averaging TTR values measured across a text in a moving-window fashion (typically with a window size of 50 words). Unlike related indices such as mean segmental TTR (MSTTR), in which the final segment of the text is ignored if it is shorter than the predetermined window size, MATTR uses all words in a text to calculate the final lexical diversity score. Only a few studies have investigated the relationship between MATTR and text length, likely because it was not available in easy-to-use text analysis tools until recently. However, a recent study (Zenker & Kyle, 2021) found that MATTR was particularly stable across L2 argumentative texts ranging from 50 to 200 words. In the realm of spoken tasks, Fergadiotis et al. (2015) found that text length did not affect MATTR scores when applied to L1 adult responses to four oral tasks. To our knowledge, however, MATTR has not been formally evaluated with L2 responses to spoken tasks. Although MATTR has been used in a few SLA studies (e.g., Hwang, 2020; Tracy-Ventura et al., 2021), it has not yet been widely adopted.

Validity of lexical diversity indices

In addition to demonstrating that a measurement tool is highly reliable, it is equally important to have clear evidence to support an argument for the validity of that tool (Chapelle et al., 2008; Kane, 2013). Studies have taken two major approaches in providing validity evidence for lexical diversity indices. In the first (and most common) approach, relationships between lexical diversity scores and proficiency scores (broadly construed) are used (Bulté & Roothoof, 2020; Engber, 1995; Jarvis, 2002; Koizumi et al., 2022; Treffers-Daller et al., 2018; Zenker & Kyle, 2021). For example, Engber (1995) investigated the relationship between lexical variety index scores (both including and excluding lexical errors) and holistic judgments of essay quality ($n = 66$). She found moderate correlations between holistic scores and lexical variety index scores both

when all words were included ($r = .450$) and when lexical errors were excluded ($r = .570$). More recently, Treffers-Daller *et al.* (2018) investigated the relationship between lexical diversity index scores for L2 essays ($n = 179$), overall Common European Framework of Reference (CEFR) proficiency levels, vocabulary scores, and writing scores. Indices that have been shown to be intrinsically related to text length (number of types, TTR, Root TTR) demonstrated significant differences between B1 and B2 levels (but not other adjacent proficiency levels) and significant moderate correlations (ranging from $r = .424$ to $.472$) were found between these indices and writing and vocabulary scores. Although indices that have been shown to be resistant to text-length effects (HD-D, vocd-D, and MTLT) did not show significant differences across adjacent CEFR levels, significant small to moderate correlations (ranging from $r = .276$ to $.344$) were found between these indices and both vocabulary scores and writing scores. In the realm of L2 spoken tasks, Bulté & Roothoof (2020) investigated the relationship between speaking proficiency scores based on an IELTS exam and various text complexity measures (including lexical diversity). The strongest correlation found was between an index that has been shown to be intrinsically related to text length (Root TTR) and speaking proficiency scores ($r = .701$). Moderate to strong correlations were found between speaking proficiency scores and two indices that have been shown to be resistant to text-length effects (HD-D, $r = 0.615$; MTLT, $r = .535$). Taken together, these results suggest that text-length stable indices of lexical diversity are indeed related to proficiency (though the effects are often moderate), providing some evidence for their validity in both written and spoken contexts. Unsurprisingly, indices affected by text length tend to be more strongly related to proficiency scores than those that effectively control for text length effects because they conflate the constructs of text length and lexical diversity.

In the second (and less common) approach, relationships between lexical diversity index scores and human judgments of lexical diversity are investigated. For example, Jarvis (2017) examined the relationship between various indices of lexical diversity and human ratings of lexical diversity in responses to a written narrative retelling task ($n = 60$). Correlations with human scores ranged from moderate to large, with the strongest ones being for MATTR ($r = .577$) and HD-D ($r = .669$). More recently, Kyle *et al.* (2021) conducted a similar study using L1 ($n = 315$) and L2 ($n = 300$) argumentative essays. They found moderate to strong correlations between human scores and HD-D ($r = .602$), MATTR ($r = .492$), and MTLT ($r = .505$). Taken together, these results provide some validity evidence for lexical diversity indices that have been shown to be at least reasonably resistant to text-length effects, though it seems clear that the psycholinguistic construct of lexical diversity consists of more than just lexical variety (see Jarvis, 2013).

Lexical diversity indices and stability across tasks

If lexical diversity scores are compared across different task prompts (or task types), as may be the case in longitudinal studies (Tracy-Ventura *et al.*, 2016) and in some cross-sectional studies (Lu, 2012; Verspoor *et al.*, 2012), it is also important to establish that lexical diversity indices are consistent from one prompt to another. Relatively few studies have systematically investigated the stability of lexical diversity indices across written task prompts or types (Alexopoulou *et al.*, 2017; Yoon, 2017; Zenker & Kyle, 2021), and none that we are aware of have done so with spoken tasks. The results of extant studies have indicated that lexical diversity scores may not be consistent across tasks. For example, Alexopoulou *et al.* (2017) found meaningful differences in MTLT scores across both task types (narrative, descriptive, professional) and prompts within

each task type using the EFCAMDAT corpus (Geertzen et al., 2014). Similarly, Zenker and Kyle (2021) found small but meaningful differences across two argumentative writing prompts in the ICNALE corpus (Ishikawa, 2011) for the nine indices of lexical diversity investigated. In contrast, Yoon (2017), who also used the ICNALE corpus, found negligible differences in *D* values across the two prompts, controlling for L1. These results preliminarily suggest that task type and task prompt may contribute to variation in lexical diversity scores (at least with respect to written tasks). It should be pointed out that there are many cases in which we might expect to see variation in lexical diversity scores across tasks as an indicator that different tasks elicit different linguistic features (Cumming et al., 2005; Kyle et al., 2016). Therefore, the observation of differences in lexical diversity scores across tasks does not necessarily reflect negatively on the reliability of an index and may in fact provide validity evidence for the inclusion of multiple production tasks in a language assessment tool. However, it is important for the degree to which task characteristics affect indices of language production such as lexical diversity to inform the design of studies of language development and/or the design of assessment tools. Therefore, further investigation is needed to determine the degree to which task type and prompt affect lexical diversity scores in spoken tasks.

Current study

The current study addresses three issues related to the stability and validity of lexical diversity indices in spoken tasks commonly used in studies of SLA by analyzing a large corpus of oral proficiency interview data. The study is guided by the following research questions:

1. What is the relationship between lexical diversity indices and text length in oral proficiency interviews?
2. To what degree are text-length stable indices of lexical diversity predictive of oral proficiency interview scores?
3. To what degree are text-length stable indices of lexical diversity stable across oral proficiency interview subtasks?

Method

Learner corpus

In this study, we used the National Institute of Information and Communications Technology Japanese Learner English (NICT JLE) corpus, which includes 1,281 transcribed oral proficiency interviews (OPIs) by Japanese learners of L2 English (Izumi et al., 2004). The version of the OPI used in corpus collection, the Standard Speaking Test (ACTFL-ALC-SST, henceforth SST), is a modified version of the American Council on the Teaching of Foreign Language (ACTFL) OPI, adjusted for the target population through the inclusion of more structured intermediate-level tasks (ACTFL-ALC Press, 1996; ALC Press, 2010; Koizumi & Hirai, 2012). The 10-to-15-minute interview consists of five stages: (1) warm-up introductions, (2) a single-picture description task, (3) a role play task, (4) a sequential picture storytelling task, and (5) wind-down questions. During the interview, the examiner “informally evaluates the test-taker’s level based on his/her responses and selects tasks appropriate to the level” (Koizumi & Hirai, 2012, p. 42). Interviews were subsequently scored by two qualified raters according to a holistic rubric (see [Appendix A](#)) that ranged from 1 to 9

points. In cases of disagreement between the two raters, a third and final rating was provided by a “master” rater (Kobayashi & Abe, 2016). The transcripts and final SST scores are publicly accessible as a part of the NICT JLE Corpus (https://alaginrc.nict.go.jp/nict_jle/index_E.html).

A Python script was used to automatically remove pauses, disfluencies (i.e., fillers, repetitions, false starts, repair), and other discourse features (e.g., Japanese words/utterances, paralinguistic cues) from the corpus using XML tags provided in the NICT JLE files. First, the interview transcripts and metadata about the interviewees and specific task types were retrieved separately from the raw corpus file. From the transcript, only the interviewees’ utterances were extracted. We then deleted disfluency features from the interviewees’ transcripts. We also deleted any utterances that were completely in Japanese (e.g., fillers, overt lexical searches). To do so, we created a list of Japanese utterances in the corpus, and a member of the research team whose L1 is Japanese determined whether each instance was an independent Japanese utterance or a Japanese phrase inserted into an English construction. Finally, we extracted the cleaned version of the text and classified it by task type (Stage 2 = single-picture description task; Stage 3 = role play task; Stage 4 = sequential picture storytelling task). Stage 1 (warm-up) and Stage 5 (wind-down) were excluded from all analyses. The Python code for this procedure can be found in the online supplementary material (<https://osf.io/ya8se>).

Table 1 shows the distributions of learners and tokens across the various proficiency levels in the corpus (including only data from Stages 2-4).

Lexical diversity indices

Lexical diversity indices were calculated with the Python version of TAALED (Kyle *et al.*, 2021; version .32). Texts were preprocessed using spaCy (Explosion AI, 2018) through the pylats package (Kyle, 2022; version .37), which was also used for post-processing. Texts were lemmatized prior to the calculation of lexical diversity scores. Any misspelled words (due to transcription errors) that did not result in the creation of an English word were ignored. An overview of each of the indices examined in this study is provided below.

Number of types

The number of types is simply a count of the number of unique lemmas in a text. Although number of types is strongly related to direct human judgements of lexical diversity (see Kyle *et al.*, 2021) and CEFR proficiency levels (Treffers-Daller *et al.*, 2018),

Table 1. Learner and token counts by proficiency level in the corpus data for Stages 2-4

Level	Learners	Tokens (Mean)	Tokens (SD)
1	3	81.67	30.07
2	35	125.54	43.90
3	222	279.04	80.16
4	482	428.03	101.99
5	236	584.32	132.55
6	130	688.23	148.82
7	77	726.13	150.55
8	56	851.54	204.72
9	40	964.90	228.76

it is also strongly related to text length and is not generally advisable to use as an index of diversity. It is included in this study as a baseline index.

Type-token ratio (TTR)

The simple type-token ratio (TTR; Johnson, 1944) is calculated as the number of unique words in the text (types) divided by the number of running words (tokens): $TTR = nTypes/nTokens$.

Root TTR

Root TTR (also known as Guiraud's index; Guiraud, 1960) is calculated as the number of types divided by the square root of the number of tokens: $Root\ TTR = nTypes/\sqrt{nTokens}$.

Maas

Maas's index (Maas, 1971) is a transformation of TTR that fits the type and token measures to a logarithmic curve: $Maas = \log(ntokens) - \log(ntypes)/\log(ntokens)^2$.

MATTR

Moving-average TTR (MATTR; Covington & McFall, 2010) is calculated by taking the moving average of TTR measurements for all segments of a given length. For MATTR with a window length of 50 tokens (MATTR 50), TTR is calculated on tokens 1–50, 2–51, 3–52, etc., and the resulting TTR values are averaged to produce the final MATTR value. Although 50-word windows are commonly used in studies on writing proficiency and development, other window lengths can be used as well (Fergadiotis et al., 2015; Tracy-Ventura et al., 2021; Treffers-Daller et al., 2022). In order to preliminarily determine an optimal window size for this study, we examined the relationship between MATTR and SST score with window sizes ranging from 1 to 100 tokens. The results of this analysis (see Figure 1) indicated that MATTR calculated with an 11-word window size resulted in the highest correlation ($r = .504$) with SST score. We therefore considered both MATTR 50 (which has been commonly used in L2 studies) and

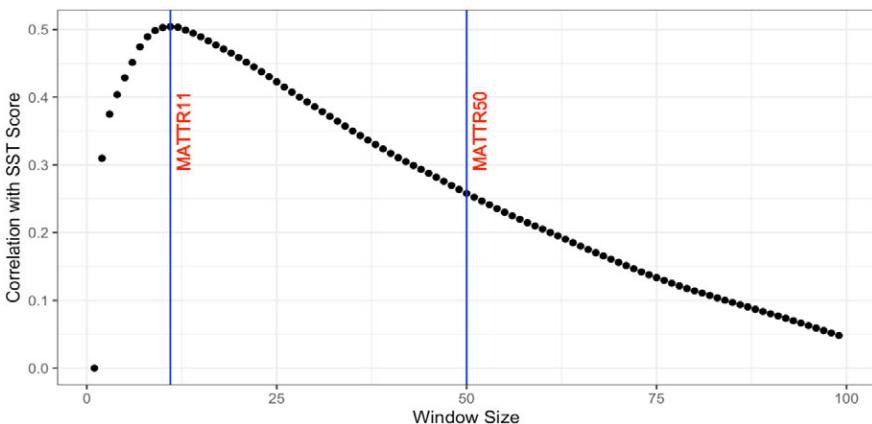


Figure 1. Optimizing MATTR window size.

MATTR 11 (which demonstrates the strongest relationship with SST score in this study) in our analyses.

HD-D

The hypergeometric distribution diversity index (HD-D; McCarthy & Jarvis, 2007) is a revised version of vocd-D that uses a hypergeometric distribution to calculate the probability of encountering a given word type in a random 42-token sample. The probabilities for each word type in the text are then added together to produce the final HD-D value. Although vocd-D and HD-D are strongly correlated (Koizumi & In'nami, 2012; McCarthy & Jarvis, 2007), McCarthy and Jarvis assert that HD-D is the value that vocd-D is attempting to approximate and that it is therefore the more appropriate choice.

MTLD

The measure of textual lexical diversity (MTLD; McCarthy, 2005; McCarthy & Jarvis, 2010) represents the average number of tokens, with a 10-token minimum, that it takes to reach a predetermined TTR cutoff value (defined as $TTR = .72$ in McCarthy & Jarvis, 2010) in a text. The resulting token lengths, which are calculated sequentially from one end of the text to the other, are referred to as factors. Partial factors occur where there are not enough remaining tokens to reach the specified TTR value. In the most commonly-used versions of MTLD, the procedure is performed both forwards and backwards on the text as a means of dealing with partial factors. Finally, the lengths of all complete factors are averaged to produce the MTLD value. In order to preliminarily determine an optimal TTR cutoff value for this study, we examined the relationship between MTLD and SST score with TTR cutoff values ranging from $TTR = .60$ to $TTR = .92$. The upper bound of the TTR cutoff values ($TTR = .92$) was set by determining the average TTR value in 10-token text segments (the default minimum MTLD factor size), as visualized in Figure 2. The results of this analysis (see Figure 3) indicated that MTLD calculated with a cutoff of $TTR = .92$ resulted in the highest correlation ($r = .430$) with SST score. We therefore considered both MTLD .72 (which has been commonly used in L2 studies) and MTLD .92 (which demonstrates the strongest relationship with SST score in this study) in our analyses.

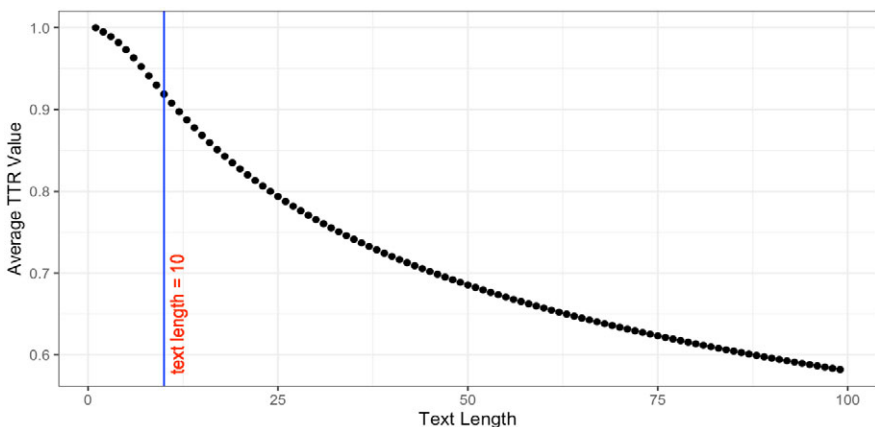


Figure 2. Relationship between average TTR value and text segment length.

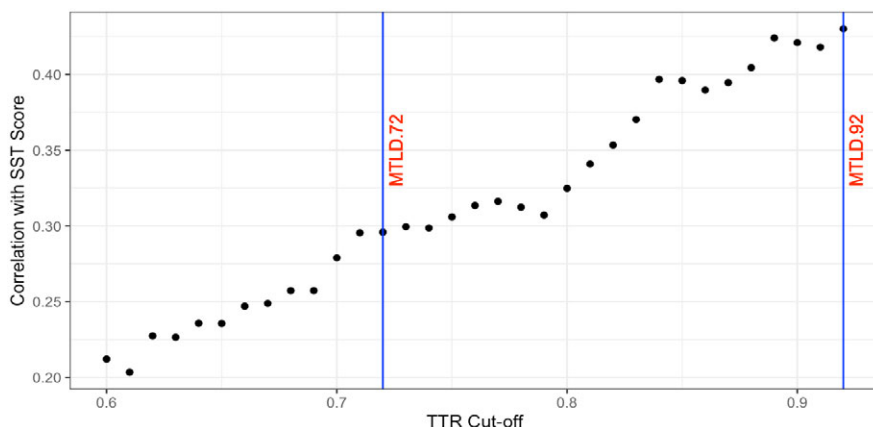


Figure 3. Optimizing TTR cut-off values for MTLTD.

Analyses

To address RQ1, which is concerned with the intrinsic relationship between lexical diversity values and text length, an analysis was conducted using the parallel sampling method (Hess et al., 1986). Text lengths from 50 to 400 tokens were examined in increments of 5 tokens. In order to standardize the parallel sampling analysis, any texts shorter than 400 words in length were excluded. In texts longer than 400 words, only the first 400 words were used in the analysis. See Table 2 for the number of responses that met the length requirements (organized by SST score).

The results were analyzed using data visualizations and with respect to correlations over the entire range (i.e., text segments of 50 to 400 words). Because previous research (Zenker & Kyle, 2021) has indicated that some indices stabilize at longer text lengths, visualizations and correlations across five equally sized text-length bins (Bin 1 = 50–115 tokens, Bin 2 = 120–185 tokens, etc.) were also examined.

To address RQ2, which is concerned with the validity of text-length stable indices of lexical diversity, correlations were conducted between the text-length stable indices and the overall proficiency scores. For this analysis, all texts were used regardless of length (see Table 1 for reference).

To address RQ3, which is concerned with the stability of lexical diversity indices across different tasks, a linear mixed-effects model was conducted for each text-length stable index. All texts were used in this analysis. In each model, the lexical diversity index was the outcome variable, Task (a three-level category) was entered as a fixed-effect, and by-participant intercepts were added as the random effect. The linear mixed-effects models were fit using the lme4 package (Bates et al., 2015). Two effect-size metrics were used to interpret the magnitude of the task effects: (a) marginal R^2 (which indicates the variance explained by the fixed effects) and conditional R^2 (which indicates the variance explained by the fixed effects + the random effects), computed through MuMIn package (Bartoń, 2019) and (b) Cohen’s d metrics for the pairwise comparisons computed using

Table 2. Number of responses per SST score level included in the parallel sampling analysis

SST score	1	2	3	4	5	6	7	8	9	Total
Number of texts	0	0	13	258	220	129	77	56	40	793

the `eff_size()` function of the `emmeans` package (Lenth *et al.*, 2019). For the interpretation of the effect size, we used Cohen's (1988) benchmark ($d > .2$ as small effects). All data and R code can be found in the online supplemental material.

Results

RQ1

To address the first research question, which was concerned with the relationship between lexical diversity values and text length, we conducted a parallel sampling analysis and a series of follow-up correlation analyses.

We present two approaches to the analysis of the results. In the first approach, the relationship between sample size and lexical diversity scores is calculated for the entire range of text lengths examined (*i.e.*, 50–400 words). Visualizations of the relationship between lexical diversity indices and text length are provided in Figure 4. Note that *z* scores are used in the visualizations to put all the lexical diversity measurements on a common scale (thus facilitating comparisons across indices). A summary of these results is also provided in Table 3. The results indicate that five of the indices are stable across the text lengths investigated, including the two versions of MTLT, the two versions of MATTR, and Maas's index. Number of types, TTR, and Root TTR all demonstrated large correlations with text length (with absolute values above $r = .800$). Finally, the lexical diversity index *D*, which in this study was operationalized as HD-D, demonstrated a moderate ($r = .505$) correlation with text length.

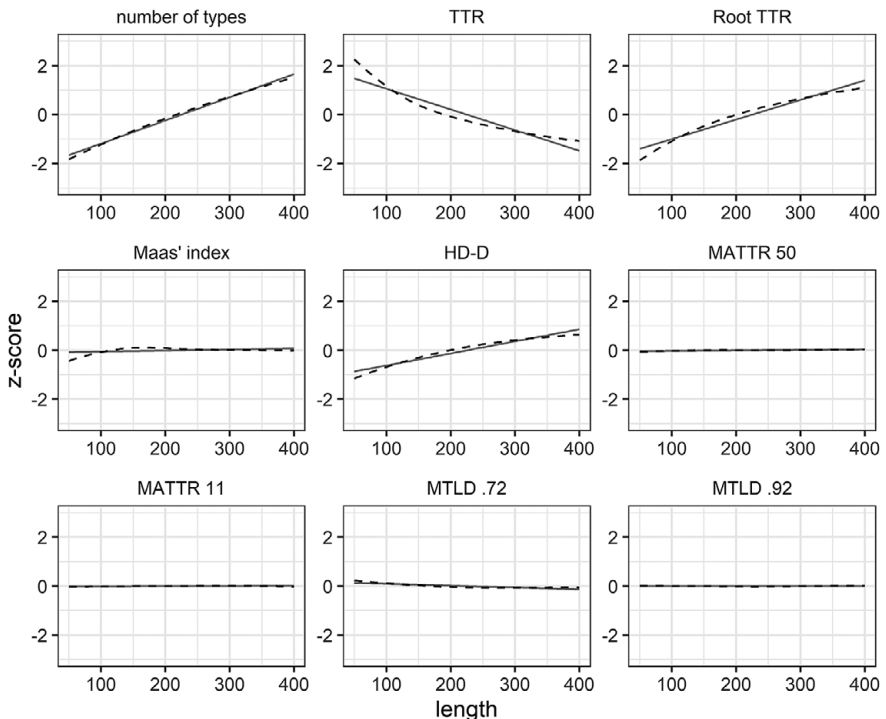


Figure 4. Relationship between lexical diversity scores and text length. Solid gray line represents the line of best fit over the entire data set. The black dashed line represents the moving line of best fit (Loess line).

Table 3. Correlations between lexical diversity indices and text length across bins

Index	Full Sample (50-400)	Bin 1 (50-115)	Bin 2 (120-185)	Bin 3 (190-255)	Bin 4 (260-325)	Bin 5 (330-400)
MTLD .92	.001	<0.001	.008	.004	.009	.003
MATTR 11	.006	.005	<.001	.007	.001	.004
MATTR 50	.021	<0.001	.016	.009	.009	.003
Maas's index	.042	.158	.009	.017	.006	.011
MTLD .72	-.076	.057	.022	.022	<.001	.002
HD-D	.505	.204	.156	.098	.076	.064
Root TTR	.821	.707	.392	.251	.175	.151
TTR	-.866	.751	.456	.298	.244	.216
number of types	.970	.947	.803	.636	.518	.465

Note. Indices are arranged in ascending order according to their correlation with segment length in the full sample (50-400 tokens).

In the second approach, we analyze the relationship between sample size and lexical diversity scores within five text-length bins (50-115; 120-185; 190-255; 260-325; 330-400) to determine whether the observed relationships change with increasing text lengths. The results of this analysis are summarized in Figure 5. The solid gray line

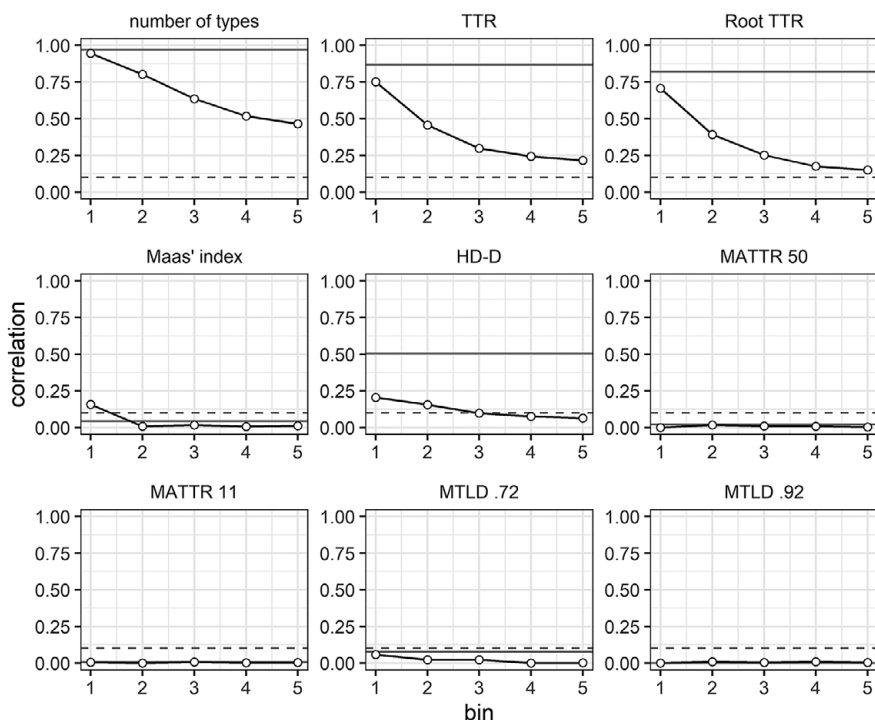


Figure 5. Relationship between lexical diversity score and text length across bins. The solid gray line represents the correlation between lexical diversity index scores and text length for texts ranging from 50 to 400 words. The dashed gray line represents the lower threshold for a “small” correlation ($r = .100$). The dots indicate the correlation between lexical diversity scores and text length within each text-length bin. Bin 1 = 50-115 words, Bin 2 = 120-185 words, Bin 3 = 190-255 words, Bin 4 = 260-325 words, and Bin 5 = 330-400 words.

indicates the correlation between lexical diversity score and text length over the entire sample. The closer this line is to zero, the more stable the lexical diversity index is across the entire dataset. The plotted points indicate the correlation between lexical diversity score and text length in a particular bin. Again, the closer these points are to zero, the more stable the index is across text lengths in that bin. The dashed line indicates the lower threshold for a “small” correlation ($r = .100$) according to Cohen (1988). A summary of these results is also provided in Table 3. The results show that the two versions of MTL D and the two versions of MATTR remain below a correlation of $r = .100$ across all bin lengths, indicating that they are particularly reliable with respect to text-length effects. Maas’s index is reliable for each bin starting at Bin 2 (120–185 words), suggesting that Maas’s index scores are likely to be comparable for texts that are at least 120 words long and do not differ by more than 65 words in length. The lexical diversity index *D*, which is operationalized as HD-D, is reliable within bins starting at Bin 3 (190–255 words). Although TTR and Root TTR become more stable as text length increases, they do not achieve full stability for any of the bins in this study.

RQ1 interim discussion

The results indicate that five of the indices examined in this study (MATTR 50, MATTR 11, MTL D .72, MTL D .92, and Maas’s index) demonstrate negligible correlations with segment length. Of these five, all indices but one (Maas’s index) are stable across the entire range of text lengths examined. In contrast to recent, large scale analyses of L2 written texts (Zenker & Kyle, 2021), the index *D* (operationalized as HD-D) demonstrated a moderate correlation with text length ($r = .505$) suggesting that it should not be used to measure lexical diversity in oral proficiency interview settings.

RQ2

To address the second research question, which was concerned with the validity of text-length stable indices of lexical diversity, a correlational analysis was conducted between five indices of lexical diversity and holistic oral proficiency interview score. For comparison, correlations between the holistic scores and both (a) number of tokens and (b) number of types are also included. Descriptive statistics are reported in Table 4, and the results of the correlation analysis are reported in Table 5.¹

Table 4. Descriptive statistics for SST and lexical diversity scores calculated across all texts in the corpus ($n = 1,281$)

Index	n	Mean	SD	Median	Min	Max
SST score	1,281	4.664	1.574	4.000	1.000	9.000
Number of tokens	1,281	495.411	224.363	461.000	51.000	1,650.000
Number of types	1,281	177.465	54.430	174.000	33.000	402.000
Maas’s index	1,281	0.059	0.005	0.059	0.024	0.078
MATTR 11	1,281	0.908	0.021	0.911	0.808	0.956
MATTR 50	1,281	0.685	0.034	0.687	0.562	0.825
MTL D .72	1,281	38.685	8.117	38.119	19.741	107.545
MTL D .92	1,281	13.074	1.191	13.016	10.015	17.500

¹Results of correlations between all indices (including those that were shown to be intrinsically affected by text length) and proficiency scores (including TTR [$r = -.681$] and Root TTR [$r = .619$]) are available in the online repository. Full correlation matrices are also available.

Table 5. Correlations between text-length stable indices and SST score

	Score	Number of tokens	Number of types	Maas's index	MATTR 11	MATTR 50	MTLD .72
Number of tokens	.831						
Number of types	.828	.959					
Maas's index	.146	.243	.049				
MATTR 11	.504	.466	.555	-.225			
MATTR 50	.258	.181	.320	-.695	.666		
MTLD .72	.296	.234	.353	-.582	.726	.862	
MTLD .92	.430	.405	.499	-.308	.814	.696	.716

RQ2 interim discussion

The results show that of the text-length stable indices, MATTR 11 demonstrates the largest correlation with SST score ($r = .504$), followed by MTLD .92 ($r = .430$). Given that these two indices provided reliable values across different text-length segments, they preliminarily represent excellent options for measuring lexical diversity in oral proficiency interview settings.

Meaningful but small correlations were also observed between SST score and Maas's index, MATTR 50, and MTLD .72, respectively. These results suggest that although Maas's index controls for the variability introduced by different token counts, it also appears to suppress the variability in lexical diversity scores that can be attributed to differences in proficiency ($r = .146$). To a lesser degree, a similar pattern is found with MATTR 50 and MTLD .72, though these indices capture more of the variability across proficiency levels.

RQ3

To address the third research question, which was concerned with the stability of lexical diversity indices across tasks within the SST, a series of linear mixed-effects models were conducted. The descriptive statistics are reported in Table 6 and the results from our analysis of differences across stages are summarized in Table 7. A full account of the model outputs is available in the online supplementary material (<https://osf.io/ya8se>).

RQ3 interim discussion

The results of the linear mixed-effects models—in particular the low R^2_{Marginal} values—indicated that the variance in lexical diversity scores explained by task (oral proficiency

Table 6. Descriptive statistics for lexical diversity scores across tasks

Index	Stage 2		Stage 3		Stage 4	
	Mean	SD	Mean	SD	Mean	SD
Number of tokens	143.948	73.977	185.376	91.837	166.086	84.981
Number of types	71.638	27.518	83.694	29.87	78.646	30.109
Maas's index	0.062	0.011	0.065	0.009	0.063	0.009
MATTR 11	0.907	0.034	0.910	0.028	0.906	0.029
MATTR 50	0.684	0.060	0.681	0.048	0.671	0.052
MTLD .72	38.171	11.435	37.684	9.441	37.869	10.303
MTLD .92	12.897	2.057	13.024	1.617	12.822	1.834

Table 7. Summary of differences across stages

Index	R^2_{Marginal}	$R^2_{\text{Conditional}}$	d (Stage 2–3)	d (Stage 2–4)	d (Stage 3–4)
Maas' index	.011	.226	–.255	–.136	.119
MATTR 11	.003	.294	–.113	.013	.127
MATTR 50	.010	.288	.053	.230	.177
MTLD .72	<.001	.254	.047	.029	–.018
MTLD .92	.002	.237	–.069	.041	.110

interview stage) was negligible. Furthermore, the difference between the $R^2_{\text{Conditional}}$ values and the R^2_{Marginal} values shows that more variance in lexical diversity scores was explained by differences across participants than by differences across tasks. A follow-up pairwise analysis using the estimated marginal means produced by the mixed-effects models supported this finding, with two minor exceptions: small differences were found between Stages 2 and 3 for Maas's index ($d = -.255$) and between Stages 2 and 4 for MATTR 50 ($d = .230$). No meaningful differences were found across tasks for MATTR 11, MTLD .72, or MTLD .92, suggesting that these indices can be used to compare lexical diversity across single-picture description, role play, and sequential picture storytelling tasks.

Discussion

In this large-scale study, we investigated aspects of the reliability and validity of indices of lexical diversity in an oral proficiency interview context. We first examined the degree to which indices of lexical diversity produced scores that were reliable (i.e., consistent) across texts of different lengths. Following previous L2 writing studies and small-scale L2 spoken studies, we found that number of types, TTR, and Root TTR are strongly and intrinsically related to text length in oral proficiency interview settings. Although these results are not surprising given previous research, it bears repeating that Root TTR conflates lexical diversity and text length given the fact that it continues to be used as an index of lexical diversity in SLA research (Bulté & Housen, 2019; Lambelet, 2021). This study also found that D is moderately and intrinsically related to text length ($r = .505$) in oral proficiency contexts. This finding contrasts with some previous studies that have focused on L1 and L2 written contexts (McCarthy & Jarvis, 2007; Zenker & Kyle, 2021) but supports previous small-scale studies that focused on oral task responses (Koizumi & In'nami, 2012). Although D (operationalized as either HD- D or vocd- D) has been used widely in SLA studies (Polat & Kim, 2014; Révész et al., 2016; Vercellotti, 2017) based on previous research that supported its use in written contexts, the findings of this study suggest that D also conflates text length and lexical diversity in oral proficiency interview settings and should not be used in these settings as a measure of lexical diversity. Following previous L1 and L2 writing studies, this study found that MTLD (both in its classic and optimized forms) was stable across oral texts ranging in length from 50 to 400 words. Similarly, MATTR (both in its classic and optimized forms) was stable across oral texts ranging in length from 50 to 400 words. Previous research in L1 and L2 contexts has shown that Maas's index is reasonably stable across text lengths (cf., Koizumi & In'nami, 2012), though usually not quite as stable as other options such as MTLD and MATTR. The findings of this study were similar—Maas's index was reasonably resistant to text-length effects in an oral proficiency interview setting, but it was less stable than MTLD and MATTR.

After determining that Maas's index, MATTR, and MTLT produced highly reliable scores across texts of different lengths, we proceeded to investigate the degree to which these indices were valid indicators of L2 spoken proficiency. The results indicated that although Maas's index was reasonably stable across different text lengths, it was only weakly related to oral proficiency scores, which calls into question its validity as a measure of L2 spoken lexical diversity. The classic versions of MATTR (MATTR 50) and MTLT (MTLT .72) demonstrated small correlations with oral proficiency scores, and the optimized versions of these indices (MATTR 11 and MTLT .92) demonstrated moderate correlations with oral proficiency scores, with MATTR 11 demonstrating the strongest relationship ($r = .504$). These results generally support previous studies that have found moderate correlations between MATTR and MTLT on the one hand and writing or speaking proficiency scores on the other (Bulté & Roothoof, 2020; Kyle et al., 2021; Treffers-Daller et al., 2018). These results suggest that MATTR and MTLT are appropriate choices for the measurement of lexical diversity in oral-proficiency interview settings and that index values obtained by averaging across shorter segments (as furnished via 11-word windows in the case of MATTR or a target TTR of .92 in the case of MTLT) are more closely related to oral proficiency ratings than those obtained by averaging across longer segments.

As a final step, we investigated the relationship between lexical diversity index scores and oral task types. To do so, we conducted linear mixed-effects models with post hoc comparisons to determine the degree to which lexical diversity scores differed across a picture description task, a role play task, and a sequential picture storytelling task. Overall, the results indicated that differences in Maas's index, MATTR, and MTLT were particularly small across these tasks. In post hoc pairwise analyses, only two meaningful (but small) differences were found: Maas's index demonstrated a small ($d = -.255$) difference between the single-picture description task and the role play task, whereas MATTR 50 demonstrated a small ($d = .230$) difference between the single-picture description task and the sequential picture storytelling task. These results diverge from those of previous L2 writing studies (Alexopoulou et al., 2017; Yoon, 2017; Zenker & Kyle, 2021) that found systematic differences across written task types and task prompts. It is possible that less variation was found across spoken task responses than in written ones because spoken texts tend to have a higher proportion of (repeated) function words, whereas written texts tend to have a higher density of content words (e.g., Biber et al., 2004; Kyle et al., 2022; Read, 2000). Function word repetition may smooth out differences in content word use across tasks. Another potential explanation for these results is that prompt differences within each task may have muted across-task differences. In the SST, much like in an ACTFL OPI, the interviewer chooses prompts based on how an interview unfolds, leading to a substantial heterogeneity in the combination of prompts that are represented in the corpus. As some previous studies (Zenker & Kyle, 2021) have found within-task (across prompt) differences in lexical diversity scores across tasks, this heterogeneity may have contributed to the observed (and substantial) random effects. More research is needed across a wider range of oral tasks to determine the degree to which lexical diversity scores are indeed stable across oral tasks.

Taken together, the results suggest that MATTR and MTLT are highly reliable and reasonably valid measures of lexical diversity in L2 oral proficiency interview settings but TTR, Root TTR, and D are not. Furthermore, the results of this study indicate that MATTR 11 is a particularly appropriate measure of lexical diversity in oral proficiency interview settings because it is stable across text lengths, demonstrates the strongest relationship with oral proficiency interview scores, and is stable across the oral tasks

investigated. In contrast, Root TTR was found to be strongly related to text length in oral proficiency interview settings, which suggests that it should not be used as an index of lexical diversity. Although some researchers (Bulté & Roothoof, 2020) have argued that Root TTR's relationship with text length is not necessarily a bad thing (e.g., because it means that a single index can be used to measure productivity and diversity simultaneously), we argue that it is important and advantageous to measure constructs in as precise a manner as is possible. For example, if researchers decide that they wish to measure proficiency using text length (e.g., as a measure of fluency or productivity) and lexical diversity, we argue that it is preferable to include them as distinct indices so that the relative contribution of each can be accounted for. In reference to the data in this study, if we use Root TTR to predict SST scores in a linear regression,² we explain more variance ($R^2 = .383$) than if we use the strongest text-length stable index (MATTR11; $R^2 = .254$), but we cannot be sure how much of the former variance is attributable to text length and how much is attributable to lexical diversity. If we include both text length and these indices in a regression, Root TTR + text length explains approximately the same amount of variance ($R^2_{\text{adj.}} = .692$) as MATTR11 + text length ($R^2_{\text{adj.}} = .707$). However, in the latter model, we can estimate that 25.4% of the variance explained by the model is attributable to lexical diversity (MATTR11) and 45.3% of the variance is attributable to text length. Such an estimation is not conceptually possible with Root TTR because it conflates lexical diversity and text length. Using indices that measure a distinct construct or subconstruct helps us understand longitudinal development and/or differences across proficiency levels more deeply and in a more fine-grained manner. Doing so also allows researchers to create composite measures that are transparent. For related arguments in support of fine-grained and precise measures of linguistic complexity, see Biber *et al.* (2011), Jarvis (2013), Kyle and Crossley (2018), and Norris & Ortega (2009), *inter alia*. Furthermore, the results indicated that *D* is also moderately related to text length ($r = .505$) and therefore should not be used as a measure of lexical diversity in oral proficiency interview settings. It should be noted that although this finding goes against conventional wisdom in SLA, both large-scale studies that have used oral L1 texts (McCarthy & Jarvis, 2007) and small-scale studies that have used oral L2 texts (Koizumi & In'nami, 2012) have found a relationship between *D* and text length.

Limitations and future directions

This study has some limitations that should be addressed in future studies. First, although our sample was large, it was also homogenous with respect to L1 (Japanese). Future research should investigate other L1 groups. Second, although the oral tasks used (single-picture description, role play, and sequential picture storytelling) are common in SLA research and L2 assessment, they do not comprehensively represent the types of oral tasks used in related studies. Future research should therefore also examine the validity and reliability of lexical diversity indices with other oral tasks (argumentative monologues, personal experience monologues, etc.). Third, we investigated the validity of text-length stable lexical diversity indices using only holistic speaking proficiency scores. Although this approach to validating lexical diversity indices is common in the field, it would also be helpful to investigate the relationship between automated lexical diversity indices and direct human judgments of the lexical diversity in oral texts following recent research in

²These analyses and all relevant data are included in the supplemental online repository.

the realm of L2 writing (Jarvis, 2017; Kyle et al., 2021). Fourth, in this study, we investigated the effects of text length and task on indices of lexical diversity, but did not investigate other potential construct-irrelevant confounds. Finally, although we investigated the stability of lexical diversity indices across the three task types used in this study, previous L2 writing research (Alexopoulou et al., 2017) has suggested that both task and prompt may affect lexical diversity scores. Therefore, in addition to examining the stability of lexical diversity scores across a wider range of tasks, future research should also probe the reliability of lexical diversity indices across different types of prompts, which was not investigated in this study.

Conclusion

Lexical diversity is and has been an important construct in the measurement of written and oral proficiency in studies of L2 development and assessment for some time. However, many indices that have been used to measure lexical diversity lack sufficient reliability and/or validity. In this study, we evaluated lexical diversity indices with respect to two aspects of reliability and one aspect of validity. The results showed that TTR, Root TTR (Guiraud's index), and *D* (operationalized as HD-*D*) demonstrated low reliability across oral task responses ranging from 50–400 words. Of the indices of lexical diversity that were highly reliable across different text lengths, MATTR and MTLD (and in particular an optimized version of MATTR) demonstrated moderate correlations with holistic oral proficiency scores, providing some validity evidence. Furthermore, the optimized versions of MATTR and MTLD were highly reliable across the three oral task types investigated. In line with previous research (primarily in written contexts), the present results suggest that SLA researchers should not use Root TTR to index lexical diversity because it conflates diversity and text length. Furthermore, the results suggest that SLA research should not use *D* as a measure of lexical diversity with spoken L2 production data because it also conflates diversity and text length. Instead, researchers should use MATTR or possibly MTLD to measure lexical diversity in L2 oral task responses given the empirical evidence that supports arguments for their reliability and validity.

Data availability statement. The experiment in this article earned Open Data and Open Materials badges for transparent practices. The materials and data are available at <https://osf.io/ya8se>.

References

- ACTFL-ALC Press. (1996). *Standard Speaking Test manual*.
 ALC Press. (2010). *The Standard Speaking Test (SST)*. <http://tsst.alc.co.jp/e/assessment.html>
 Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, 67, 180–208.
 Barton, K. (2019). *MuMIn: Multi-Model Inference* (1.43.6) [Computer software]. <https://cran.r-project.org/web/packages/MuMIn/index.html>
 Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48.
 Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 Spoken and Written Academic Language corpus*. TOEFL Monograph Series.
 Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45, 5–35.

- Bulté, B., & Housen, A. (2019). Beginning L2 complexity development in CLIL and non-CLIL secondary education. *Instructed Second Language Acquisition*, 3, 153–180.
- Bulté, B., & Roothoof, H. (2020). Investigating the interrelationship between rated L2 proficiency and linguistic complexity in L2 speech. *System*, 91, Article 102246.
- Carlson, S. B., Bridgeman, B., Camp, R., & Waanders, J. (1985). Relationship of admission test scores to writing performance of native and nonnative speakers of English. *ETS Research Report Series*, 1985, i–137.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.
- Chotlos, J. W. (1944). Studies in language behavior IV: A statistical and comparative analysis of individual written language samples. *Psychological Monographs*, 56, 75–111.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.
- Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17, 94–100.
- Cumming, A. H., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, 10, 5–43.
- Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4, 139–155.
- Explosion AI. (2018). *SpaCy language models*. https://spacy.io/models/en#en_core_web_sm
- Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research*, 58, 840–852.
- Geertzen, J., Alexopoulou, T., & Korhonen, A. (2014). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In R. T. Millar, K. I. Martin, C. M. Eddington, N. M. Henery, & A. Tseng (Eds.), *Selected proceedings of the 31st Second Language Research Forum* (pp. 240–254). Cascadilla Proceedings Project.
- Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique [Problems and methods of linguistic statistics]*. Reidel.
- Hess, C. W., Sefton, K. M., & Landry, R. G. (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech, Language, and Hearing Research*, 29, 129–134.
- Hwang, H. (2020). *A contrast between VP-Ellipsis and Gapping in English: L1 acquisition, L2 acquisition, and L2 processing* [Unpublished doctoral dissertation]. University of Hawai'i at Mānoa.
- Ishikawa, S. (2011). A new horizon in learner corpus studies: The aim of the ICNALE project. In G. Weir, S. Ishikawa, & K. Poonpon (Eds.), *Corpora and language technologies in teaching, learning and research* (pp. 3–11). University of Strathclyde Press.
- Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29, 24–49.
- Izumi, E., Uchimoto, K., & Isahara, H. (2004). The NICT JLE Corpus: Exploiting the language learners' speech database for research and education. *International Journal of The Computer, the Internet and Management*, 12, 119–125.
- Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19, 57–84.
- Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, 63, 87–106.
- Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, 34, 537–553.
- Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, 12, 377–403.
- Johnson, W. (1944). Studies in language behavior I: A program of research. *Psychological Monographs*, 56, 1–15. <https://doi.org/10.1037/h0093508>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73.
- Kobayashi, Y., & Abe, M. (2016). Automated scoring of L2 spoken English with random forests. *Journal of Pan-Pacific Association of Applied Linguistics*, 20, 55–73.
- Koizumi, R., & Hirai, A. (2012). Comparing the story retelling speaking test with other speaking tests. *JALT Journal*, 34, 35–60.

- Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, 40, 554–564.
- Koizumi, R., In'nami, Y., & Jeon, E. H. (2022). L2 speaking and its internal correlates: A meta-analysis. In E. H. Jeon & Y. In'nami (Eds.), *Understanding L2 proficiency: Theoretical and meta-analytic investigations* (pp. 307–338). John Benjamins.
- Kyle, K. (2022). *PyLats Python package* (.37) [Python]. <https://pypi.org/project/pylats/>
- Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102, 333–349.
- Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, 18, 154–170.
- Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, 33, 319–340.
- Kyle, K., Eguchi, M., Choe, A. T., & LaFlair, G. (2022). Register variation in spoken and written language use across technology-mediated and non-technology-mediated learning environments. *Language Testing*, 39, 618–648.
- Lambelet, A. (2021). Lexical diversity development in newly arrived parent-child immigrant pairs: Aptitude, age, exposure, and anxiety. *Annual Review of Applied Linguistics*, 41, 76–94.
- Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25–42). University of Michigan Press.
- Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). *R package Emmeans: Estimated marginal means, AKA least-squares means* (1.47) [Computer software]. <https://github.com/rvleenth/emmeans>
- Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96, 190–208.
- Maas, H. D. (1971). Über den Zusammenhang zwischen Wortschatzumfang und Länge eines Textes [On the connection between vocabulary breadth and text length]. *Zeitschrift Für Literaturwissenschaft Und Linguistik*, 2, 73–96.
- MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed., Vol. 2). Lawrence Erlbaum Associates.
- Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language* (Vol. 12, pp. 58–71). Multilingual Matters.
- Malvern, D. D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Palgrave Macmillan.
- McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)* [Doctoral dissertation, The University of Memphis].
- McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, 24, 459–488.
- McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42, 381–392.
- Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30, 555–578.
- Pfenniger, S. (2020). The dynamic multicausality of age of first bilingual language exposure: Evidence from a longitudinal content and language integrated learning study with dense time serial measurements. *The Modern Language Journal*, 104, 662–686.
- Polat, B., & Kim, Y. (2014). Dynamics of complexity and accuracy: A longitudinal case study of advanced untutored development. *Applied Linguistics*, 35, 184–207.
- Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.
- Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, 37, 828–848.
- Tracy-Ventura, N., Huensch, A., & Mitchell, R. (2021). Understanding the long-term evolution of L2 lexical diversity: The contribution of a longitudinal learner corpus. In B. Le Bruyn & M. Paquot (Eds.), *Learner corpus research meets second language acquisition* (pp. 148–171). Cambridge University Press.
- Tracy-Ventura, N., Mitchell, R., & McManus, K. (2016). The LANGSNAP longitudinal learner corpus. In M. Alonso-Ramos (Ed.), *Spanish learner corpus research: Current trends and future perspectives* (Vol. 78, pp. 117–142). John Benjamins.

- Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79–104). John Benjamins.
- Treffers-Daller, J., Mukhopadhyay, L., Balasubramanian, A., Tamboli, V., & Tsimpli, I. (2022). How ready are Indian primary school children for English medium instruction? An analysis of the relationship between the reading skills of low-SES children, their oral vocabulary and English input in the classroom in government schools in India. *Applied Linguistics*, 43, 746–775.
- Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, 39, 302–327.
- Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32, 323–352.
- Vercellotti, M. L. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, 38, 90–111.
- Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, 21, 239–263.
- Vidal, K., & Jarvis, S. (2020). Effects of English-medium instruction on Spanish students' proficiency and lexical diversity in English. *Language Teaching Research*, 24, 568–587.
- Yoon, H.-J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System*, 66, 130–141.
- Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, 47, Article 100505.

Appendix A

The SST Rubric	
Score	Descriptors
9	A Level-9 speaker can proficiently respond to any topics ranging from familiar ones to those of general interest. He/she can comfortably speak in any tense, for example, to narrate and describe and can effectively deal with unexpected complications as well. In addition, a speaker at this level can construct his/her response in a logical paragraph-like structure. Though few unconsciously made minor errors in grammar and word choices may be present, such do not impede comprehension at all.
8	A Level-8 speaker can proficiently respond to various topics ranging from familiar ones to those of general interest. He/she is able to deal with unexpected complications most of the time. Though rare, flaws in grammar are still present. Tense control may still weaken in certain cases, and the speaker may have some difficulty in complex sentence construction. The responses are mostly organized but sometimes lack fluency and/or may include minor word choice errors; needless to say, they do not have a significant impact on listeners' comprehension.
7	A Level-7 speaker can communicate with proficiency necessary to live and survive in English-speaking countries. He/she is able to deal with complicated situations as well, but effort is required in doing so as grammar/fluency control and speech organization may weaken. Nonetheless, a speaker at this level has noticeable strengths supporting their proficiency such as abundant volume or native-like pronunciation.
6	A Level-6 speaker can communicate with proficiency necessary to live and survive in English-speaking countries. The speaker can somewhat effortlessly string simple sentences together to express his/her thoughts; however, as the sentences become longer and more complex, fluency and grammar control sometimes weaken. Tense control errors may still often be present. Pronunciation varies from speaker to speaker. Some may sound native-like whereas others are still influenced by their native language.
5	A Level-5 speaker can maintain simple communication by talking about familiar topics, answering and asking simple questions. The speaker can also add extra information and details to his/her responses, but as sentences become longer and more complex, accuracy

(Continued)

(Continued)

The SST Rubric	
Score	Descriptors
	weakens. For example, the speaker's grammar control and fluency may weaken, and/or it may require much time for the speaker to complete them. Word choices and pronunciation are still influenced by the speaker's native language; however, listeners used to non-native English speakers would not have trouble understanding the responses.
4	A Level-4 speaker can maintain simple communication by talking about familiar topics and asking simple questions. A speaker at this level can connect simple short sentences to convey his/her thoughts, but fluency is disturbed doing so. With effort, the speaker can manage to respond to what has been asked, but he/she still cannot actively interact. The speaker's pronunciation and word choices may still be influenced by his/her native language, but the impact is insignificant and listeners used to nonnative English speakers would not have trouble understanding him/her.
3	In addition to memorized set phrases, a Level-3 speaker, at times, creates simple short sentences to convey his/her thoughts. However, the speaker is only able to do so when the content of the response is very familiar to him/her, and major errors in grammar and word choices impeding comprehension are still present. Since a great amount of effort is required to create, the responses are often slow, thus requiring listeners' patience. In addition, the pronunciation of a speaker at this level is still influenced by his/her native language and is, at times, difficult to understand without clarification.
2	With a great amount of effort, a Level-2 speaker may provide the bare minimum information necessary to maintain communication when answering simple questions regarding his/her everyday life. However, the responses are mainly just a combination of words, phrases, and memorized set expressions. There are long pauses in the responses, and in some cases, we may hear the speaker simply repeat what was heard in the question. The speaker may attempt to create in sentences; however, major errors in grammar and word choices are frequent. Even listeners who are used to hearing nonnative English speakers have difficulty understanding a speaker at this level.
1	A Level-1 speaker cannot communicate in English. The speaker may identify him/herself and make simple greetings using memorized phrases. However, in most cases, the speaker can only speak in fragments of sentences, basically just listing simple vocabulary such as numbers, days of the week, colors, and so on. He/she can rarely respond to questions, and even when showing some sort of response, it takes a tremendous amount of time doing so. In addition, the pronunciation of a speaker at this level is heavily influenced by his/her native language making it significantly difficult to understand the response.

Note. The original rubric was accessed at <http://tsst.alc.co.jp/sst/e/index.html> on January 22, 2020. The rubric (with exemplars from each level) can now be found at <https://tsst.alc.co.jp/biz/en/level/>.

Cite this article: Kyle, K., Sung, H., Eguchi, M., & Zenker, F. (2024). Evaluating evidence for the reliability and validity of lexical diversity indices in L2 oral task responses. *Studies in Second Language Acquisition*, 46: 278–299. <https://doi.org/10.1017/S0272263123000402>