OPEN ACCESS

Vocabulary Learning and Instruction
ISSN 2981-9954
Volume 15 (2026)
https://doi.org/10.29140/vli.2026.103233

Castledown

Willingness to Communicate and Lexical Complexity: Exploring Cognitive Engagement During Tasks

Paul Leeming
Kindai University – Japan
paulleeming@yahoo.com

Joseph P. Vitta
(corresponding author)
Waseda University – Japan
vittajp@waseda.jp

Abstract

Willingness to communicate (WTC) is an individual difference variable that is argued to predict the likelihood of a person engaging in communication when the opportunity arises. Researchers have investigated the predictors of WTC, and more recently have looked at how it relates to the amount of language that students use when engaged in task-based discussions (Leeming et al., 2024). This engagement in interaction is hypothesized to lead to language learning. While the amount of language produced could be considered to be a measure of behavioral engagement, the quality of language produced is also important. Students may be talking at length, but not actually contributing meaningfully to the discussion. This paper describes an exploratory study that attempted to answer a simple question: When students are highly willing to communicate, do they use a greater complexity of lexis? This complexity of vocabulary was taken to represent cognitive engagement with the task. Students from two universities completed a questionnaire measuring WTC, and then completed an eight-minute quasi-academic discussion task in groups of three. Their output was transcribed and analyzed for linguistic complexity, which was then correlated with their WTC. Results showed that students who were higher in WTC produced language that

had greater lexical diversity, but not more sophisticated lexis. This suggests that complexity of vocabulary may be a useful variable to consider when examining the relationship between WTC and task engagement, providing additional information beyond behavioral engagement.

Keywords: willingness to communicate, cognitive engagement, psychology in language learning.

Willingness to communicate (WTC) has garnered significant interest among language learning researchers, perhaps due to its direct relevance to performance in the language classroom. Most teachers have experienced situations where learners seem competent in the L2, motivated to improve their proficiency, and yet are reluctant to take advantage of opportunities to interact with their classmates. WTC is an individual difference (ID) variable that is hypothesized to represent the chances that a person will take advantage of opportunities to communicate with others (MacIntyre et al., 1998). Put simply, when given a chance for interaction, will a student use it and speak in the L2, or will they remain silent? Research has modeled possible predictors of WTC in an attempt to show variables that may influence the level of WTC that a learner has (see Teimouri et al., 2022, for an example examining the relationship between grit and WTC), but more recently there has been interest in determining the impact of WTC on language use. Leeming et al. (2024) conducted a study to examine the relationship between several variables including WTC, and the amount of language produced by learners during an academic discussion task. They found that WTC was a positive predictor of the quantity of language output, but their focus was on establishing an empirical link between WTC and language production, and therefore their study did not investigate the quality of the language produced. Quantity of language output can be used to represent behavioral engagement with a task (Leeming, 2019). This exploratory study aimed to build on the work of Leeming et al. (2024) by examining their data set, but with a view to quality of language as measured by several counts of lexical sophistication. The quality of language is likely to represent cognitive engagement, and is important, as previous research has shown that students may fake behavioral engagement by producing language for the benefit of the observer, with limited concern regarding what they are actually saying (Mercer et al., 2021). Theoretically, if students are high in WTC, they should not only produce more language (behavioral engagement), but should also use a wider range of lexis in order to express themselves more effectively (cognitive engagement).

WTC, Task-Based Language Teaching, and Engagement

WTC originated as a variable in L1 research but has become established as one of the more popular variables in language learning research (Dörnyei & Ryan, 2015). The model of WTC introduced by MacIntyre et al. (1998) shows how WTC is a complex

variable predicted by a variety of factors at different levels, including the social context of the interaction, the motivation of the learner, and also their desire to speak to a given interlocutor. Ultimately, at the top of the pyramid model is L2 use, and this relationship was established empirically by Leeming et al. (2024) who showed that general WTC predicts the number of words produced in a quasi-academic group discussion task.

These kinds of discussion tasks are often part of a task-based approach to language learning. Task-based language teaching (TBLT) is argued to benefit learners as it provides opportunities for meaningful communication in the classroom (Ellis et al., 2020). A language task will have some kind of gap (often an opinion gap or information gap), requiring students to communicate to successfully complete the task. In contrast to traditional language exercises which focus on accurate use of pre-taught grammar structures, TBLT more closely mirrors real world communication. Although typically interested in how task-related factors such as planning time, or task complexity influence the language learning potential of tasks, researchers are turning their focus to learners, and considering ID variables that may be of importance when students work together.

A variable that is garnering increasing interest in the field of TBLT is task engagement (Hiver & Wu, 2023). Engagement is increasingly used as an outcome variable in TBLT research, as more than motivation, it represents what students actually do when completing tasks (Hiver et al., 2024; Leeming & Harris, 2025). The amount of language produced could be considered to be a proxy for behavioral engagement (Hiver & Wu, 2023). However, behavioral engagement does have limitations, as students can easily feign involvement by talking at length, without contributing to the task, or communicating ideas to other learners (Mercer et al., 2021). As a result, researchers are interested in measuring not only behavioral, but also other forms of engagement such as cognitive and affective engagement (Hiver & Wu, 2023). Cognitive engagement shows deeper processing and deliberate attention to the language, and can be measured in terms of the number of times that students ask follow-up questions, or respond to something said by another speaker. In studies using discussion tasks, cognitive engagement is commonly measured using instances of language related episodes where students take time to explicitly discuss problems with the language (Leeming, 2019). Affective engagement represents feelings of enjoyment while undertaking a given task, and is generally measured by questionnaires following task completion (Hiver & Wu, 2023).

Another element of speech that could be taken to represent cognitive engagement is the complexity of lexis used during a task. If a learner has high general WTC and is making a concerted effort to communicate, it is reasonable to hypothesize that they will use a wider range of vocabulary than a learner who is not interested in, or willing to communicate. To the best of the authors' knowledge, there is no empirical research that has investigated this relationship.

Lexical Complexity in TBLT

The complexity-accuracy-fluency (CAF) paradigm of understanding language ability (see Figure 1) is the starting point to define lexical complexity (Pallotti, 2009).

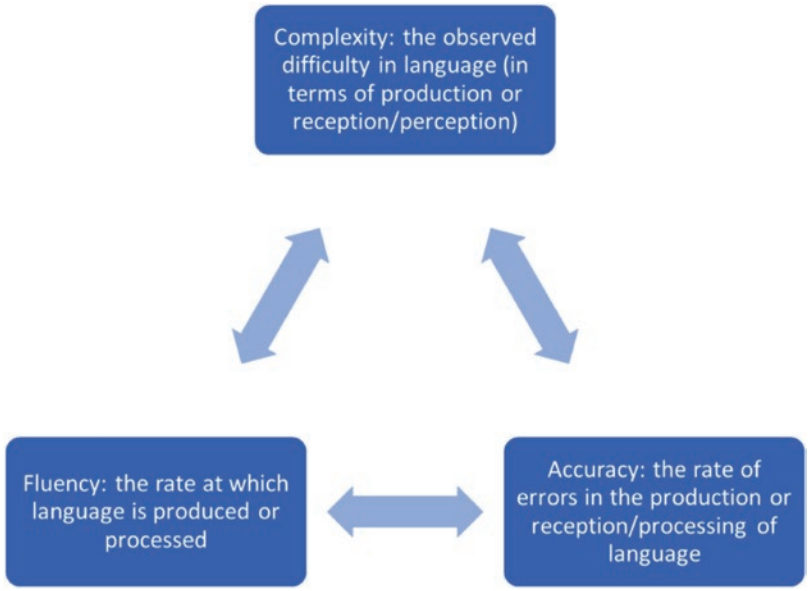

Figure 1 *The CAF Model.*

Complexity in the CAF model has been debated in relation to its constituent components, but linguistic complexity has consistently been viewed as one of the core domains under the complexity umbrella (see Figure 2). As with most areas in the CAF model, linguistic complexity's underlying constructs have been debated, but a two-factor conceptualization (lexical and syntactic complexity) is defensible in the literature (see Bultė & Housen, 2014; Lu, 2011, 2012) as concepts such as (bound) morphology and n-gram frequency can be subsumed under either term (see arguments in Vitta, 2019).

As reviewed by Vitta et al. (2023), lexical complexity is the extent to which a text's lexical features, either single words or multiword constructions (e.g., n-grams or formulaic expressions), make the text difficult to either produce (productive proficiency) or understand (receptive proficiency). As tasks and corresponding text types become more difficult, they will tend to require more complex lexical (and syntactic) structures. A journal article, for instance, that has 10% of its content words in the 8k lemma band of the *Corpus of Contemporary American English* (COCA; Davies, 2008–) is more lexically complex than a children's book where all content words are within the 1–2k bands. The components of lexical complexity have been debated but the relevant literature defends a density-diversity-sophistication conceptualization (see Figure 3) of lexical complexity's sub-domains (Lu, 2012; Vitta et al., 2023). Lexical density and diversity are somewhat straightforward in their measurement where the latter is often corrected to account for variations in text length (Kyle et al., 2021). Regarding lexical sophistication, there is robust debate on its operationalization which is outside of the focus of this study. Following the approach of Kim and Crossley (2018) and Vitta et al. (2023), the *Tool for the Automatic Analysis of Lexical Sophistication* (TAALES;

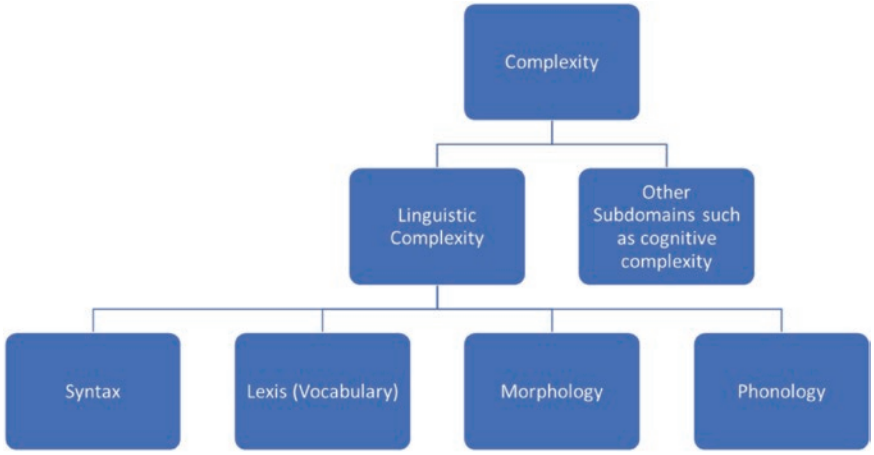

Figure 2 *Situating Linguistic Complexity within the 'C' in CAF (adapted from Bulté & Housen, 2012).*

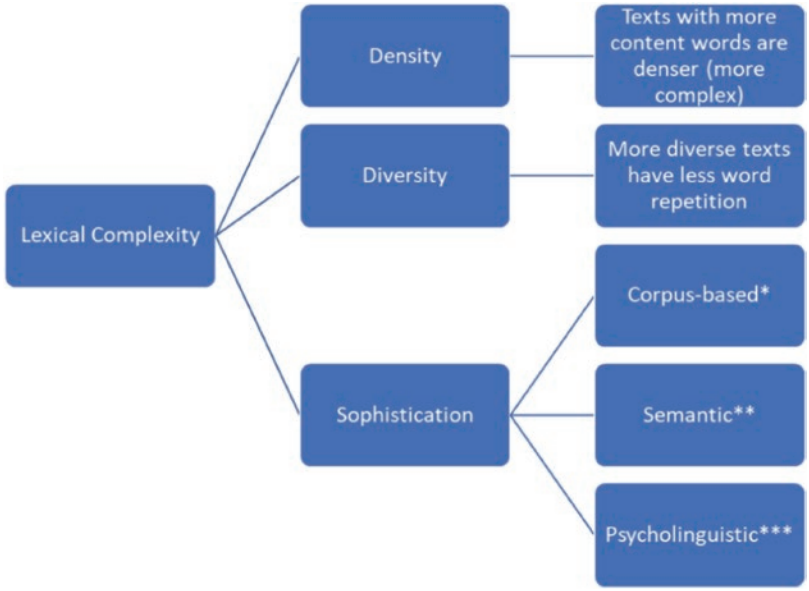

Figure 3 *The Sub-Domains of Lexical Complexity.*

Note: *Corpus-based lexical sophistication is determined by the frequency of a lexical item's (single word and multiword construction) use (Cobb, 2013); **Semantic-based lexical sophistication is determined by a lexical item's (usually measured at the single word level) semantic properties, e.g., how many different meanings does a word have (polysemy – Fellbaum, 1998); ***Average age of acquisition of a word would be an example of a psycholinguistic lexical sophistication measurement, e.g., hug (M = 2.58 years' old); prison (M = 7.50) (Kuperman et al., 2012).

Kyle et al., 2018) guiding literature is employed as a quasi-theoretical defense for measuring lexical sophistication in this current study (see Table 1 for the selected subconstructs of lexical sophistication selected from Kyle et al., 2018). Unlike Kim and Crossley, however, we excluded newer areas of measurement (relative to the first version of TAALES; Kyle et al., 2018) such as lexical decision metrics as they share conceptual overlap with psycholinguists and we did not wish to "overload" the number of tests we conducted.

The Current Study

This exploratory study builds on prior research that established the relationship between WTC and behavioral engagement in a task as measured by the number of words produced (Leeming et al., 2024). Using the data set from their study, we investigated the relationship between WTC and lexical complexity which served as a proxy for cognitive engagement. The research question driving this study was:

> What is the relationship between WTC and cognitive task engagement represented as lexical complexity?

We presupposed ("hypothesize" has been purposefully avoided given the exploratory nature of this study) that students higher in WTC would be more motivated to communicate their message to the group, and therefore would demonstrate more complex lexis in their spoken output. As with Leeming et al. (2024), WTC was conceptualized as the predictor variable and cognitive task engagement was set as the outcome variable. Given the cross-sectional nature of the dataset and the exploratory nature of the study, however, we do not make claims of causality per se. Rather, we are exploring the ability of WTC to antecede cognitive engagement as measured by the observed lexical complexity in L2 spoken task output. Trait and state WTC can coexist and are hypothesized to both influence student engagement with tasks. Although there is increasing interest in task-related WTC, Leeming et al. (2024) measured general WTC at the trait level, and their data is used in the current study.

Methodology

Participants and Data Collection

As this paper reanalyzes data from Leeming et al. (2024), abridged reporting is presented. General WTC was measured using Likert-scale survey items (from Weaver, 2010 as cited in Leeming et al., 2024; for survey items, data, and other supplementary materials see the Open Science Framework at https://osf.io/v6c32) administered via Google Forms, translated into Japanese, and initially validated through back-translation. Survey items asked students to rate their WTC in a number of general communicative situations. The testing sample (N = 262, from two Japanese university sites) first completed the WTC survey before engaging in three synchronous, dialogic, divergent group discussion tasks over four weeks (three persons to a group; for detailed description of task; see Leeming et al., 2024). In the final week, students recorded an 8-minute discussion aligned with CEFR B1/B2 descriptors

Table 1 *The Current Study's Operationalization of Lexical Complexity*

Lexical Complexity Subdomain	Measure(s) Employed	Tool(s) Employed to Derive Measures	Total Number of Measures
Lexical density	– Content words in text ÷ total words in text	TAALES output with manual computation in Excel	1
Lexical diversity	– Guiraud's (1954) index[1]: unique tokens/words in text ÷ the square root of total tokens/ words in text	*Tool for the Automatic analysis of Lexical Diversity* (Kyle et al., 2021)	1
Lexical sophistication: Single word frequency	– COCA spoken frequency content words (logged)[2] – SUBTLEXus content words (logged)[3]	TAALES	2
Lexical sophistication: Bigram frequency	– COCA spoken bigram frequency (logged)[4] – COCA bigram portion 30k-to-100k[5]	TAALES	9
Lexical sophistication: Trigram frequency	– COCA spoken trigram frequency (logged) – COCA trigram portion 30k-to-100k	TAALES	9
Lexical sophistication: Semantic Network[6]	– polysemy: content words – hypernymy: verbs (all pathways averaged) – hypernymy: nouns (all pathways averaged)	TAALES	3
Lexical Sophistication: Psycholinguistic properties	Age of Acquisition (content words)[7]	TAALES	1

[1]This measure of lexical sophistication helps to correct for variation in text length while preserving the analysis of the entire text.

[2]As the modality was speaking, the spoken corpus of COCA was used and logged to correct for Zipfan distribution (applied to all frequency measures). Precedent suggests that the measure works best when focused on content words (see e.g., Kyle & Crossley, 2015, 2016; Vitta et al., 2023).

[3]SUBTELXus was employed for robustness as it is derived from American media (e.g., movies) subtitles (Brysbaert & New, 2009).

[4]Bigrams are two-word combinations (e.g., I know) that occur to a greater degree than chance would predict. Trigrams are three-word combinations in the same manner.

[5]A portion measure uses k-level thresholds. For example, the 30k COCA bigram portion measure is the proportion of bigrams within the analyzed text that are among the most frequent 30k bigrams.

[6]Polysemy measures the different senses a word (lemma) has. Hypernymy, as operationalized in the TAALES, is the extent to which a word fits into broader superordinate categories. 'Gorilla' therefore has greater hypernymy than 'animal.' TAALES derives these values from WordNet (Fellbaum, 1998).

[7]Age of acquisition (Kuperman et al., 2012) alone is employed as a proxy for this conceptual area.

(see Supplementary Material of Leeming et al., 2024). Spoken output was transcribed. The task itself involved students discussing their preferences for living in either the countryside or city with preceding pre-tasks to help them prepare content. It was "folded" into the curriculum at each site where students studied English on a weekly basis. In the preceding two lessons (e.g., the two weeks before the recorded task), participants engaged in similar discussion tasks to get acquainted with the task procedure. The task(s) was presented in Leeming et al. (2024) as being weak (i.e., a discussion mimics a "real life" experience, but there was also a clear learning outcome present), synchronous (i.e., students interacted in real time), and divergent (i.e., students did not need to reach consensus or work towards a common goal—instead, they worked to share their own viewpoints). Another design feature of Leeming et al. (2024) worth mentioning here is that participants completed the WTC survey (along with other ID questionnaires) three weeks before the recorded task. This was to avoid priming and biasing effects while still being temporally adjacent to the recorded task.

After exclusions, 230 participants' transcribed spoken texts remained for lexical complexity analysis (thus $N = 230$ for the current study). Although English speaking proficiency data were not available for individual participants, Leeming et al. (2024) showed that the language proficiency for the two sites was roughly equivalent where the majority of participants at both sites were B1 on the CEFR.

Data Analysis: Instrumentation and Testing Procedure

WTC

Leeming et al. (2024) measured general WTC and presented a split-sample/cross-validation process (exploratory factor analysis [EFA] testing sample: $N = 209$) where the resulting confirmatory factor analysis (CFA) factor solution included 14 of the 16 proposed indicators for the construct (items 6 and 12 removed). As we intended to use parametric testing, the 14 remaining WTC z-scores were aggregated with strong underlying psychometric justification where principal components analysis suggested only one significant factor via an observed elbow distribution where the first factor accounted for 50% of the variance and its eigenvalue was approximately 7× the second factor's (for PCA readouts; see Supplementary Material B). PCA was performed using JASP 0.19.3 (JASP Team, 2025); SPSS ver. 30 and Excel were used for other statistical analyses. With unidimensionality observed (i.e., an empirical justification for construct validity), internal reliability was assessed and found to be strong (Cronbach's $\alpha = .92$).

Lexical Complexity

Lexical complexity was measured using 26 metrics subsumed under three areas using the theoretical justifications found in the preceding text. Table 1 presents the exact measures used and the tools from which they were derived:

All employed lexical complexity metrics had their bivariate associations checked with the word count outcome variable from Leeming et al. (2024). No observed absolute r value (maximum: lexical diversity [$r = .56$]; minimum: hypernymy_verb [$r = .07$]) was above the conventional collinearity thresholds of .70 to .90 found in the literature

(see discussion in Dormann et al., 2013). This observation supported the notion that the current study was exploring different conceptual areas than Leeming et al. (2024).

Power Analysis

Despite the exploratory nature of this study, a sensitivity power analysis was conducted to avoid the pitfall of using statistical significance alone to identify meaningful findings. In a sensitivity power analysis, the number of participants (in this case, $N = 230$) is referenced along with acceptable thresholds for false positives (in this case, .01 to account for the multiple tests conducted in this study—admittedly not a full correction) and false negatives (.20 following convention). Using G*Power (Faul et al., 2007), it was determined that $r = |.22|$ was the minimum meaningful effect size. This value corresponded to meta-analytic reviews of educational research (see Hattie, 2009; Vitta et al., 2022).

Testing Procedure

The testing procedure reflected the study's exploratory nature in that bivariate predictive (WTC: predictor) associations, using Pearson's r (which is identical to simple regression when assigning predictor/outcome variable status) were analyzed with the measured lexical complexity as outcome variables. Multivariate models (i.e., multiple outcome lexical complexity variables in the same model) were avoided as the intent of this current study was just to signal potential predictive associations of interest. Guided by the preceding sensitivity power analysis, a noteworthy association was identified using the following thresholds: $r > .22$, $p \leq .01$. As with Leeming et al. (2024), the potential biasing influence of site ($k = 2$) was checked as well as the assumptions of parametric testing (see Vitta et al., 2022, 2023).

Results

For the sake of parsimony, the full correlation matrix table is presented in Supplementary Material C. Of the 26 correlations analyzed, the association between WTC and lexical diversity, $r = .38$, $p = .001$, was the only noteworthy association (all assumptions of parametric testing were checked and met; see OSF). Site's association, $r = .17$, $p = .009$ with the outcome variable was below the $r > .22$ threshold. Given that p value was below the chosen alpha level, a post-hoc semi-partial correlation was conducted with WTC as the focal predictor (and site as X2). Semi-partial correlations are useful as a metric in this regard (see Abdi, 2007; Vitta et al., 2025) because only the "unique" variance (i.e., not shared between site and WTC) is regressed onto the outcome variable (lexical diversity). The observed semi-partial correlation between WTC and lexical diversity, $r_{sp} = .34$, $p < .001$, was nearly identical to its bivariate association and thus concerns about the potential bias of site were mitigated.

Discussion

Although Leeming et al. (2024) were able to show a significant relationship between general WTC and the number of words produced (behavioral engagement) by students involved in a quasi-academic discussion task, their study did not measure the

quality of language produced in terms of the lexis used (cognitive engagement). In this exploratory research, we reanalyzed the data from that study to determine the relationship between WTC and cognitive engagement as represented by lexical complexity. In order to produce complex language, learners cannot simply be going through the motions when speaking, but must be cognitively engaged in order to effectively draw on their knowledge of lexis to communicate their ideas.

The results of our study showed that the only significant relationship between WTC and the 26 different measures of lexical complexity was with lexical diversity. This corresponds to our presupposition, as learners who are more willing to communicate show greater cognitive engagement in the task, and as a result use a wider variety of language to express their ideas to their interlocutors. This helps to support the conclusions of Leeming et al. (2024) that WTC predicts behavior during discussion tasks. Students are not only more behaviorally engaged with the task, but are also cognitively engaged, attempting to convey their message. WTC relates to the level of risk that they are willing to take in using a variety of lexis.

The range of language used by a student during a task is likely to vary depending on how willing they are to cognitively engage with the discussion and the other members. While this may be influenced by task-specific factors, trait-level WTC was found to be important in this study. When WTC is low, a student may simply use repetitive expressions, or short one-word responses. Conversely, when WTC is high, it is likely that a student will work hard to express themselves and use the full range of lexis that is at their disposal. This provides a possible explanation for the link between WTC and lexical diversity in the current study. Lexical diversity may present a possible measure of engagement in TBLT in addition to other measures such as LREs or use of the idiodynamic method (Sato et al., 2025).

The other measures of lexical complexity were not related to WTC. Lexical density has been posited as relating to the information contained in speech (Lu, 2012) and is more likely to be constrained by the nature of the task, especially speaking where users may employ less function words (e.g., prepositions and articles) than writing. It is possible that ID variables such as WTC have less bearing on lexical density than the formality and modality of the task. In terms of measures of lexical sophistication, WTC failed to significantly relate to any of the variables used. Students high in WTC will be keen to get their point across and be understood by their interlocutor, and advanced lexis is unlikely to facilitate smooth communication among B1 level students engaged in a discussion task. Additionally, less sophisticated lexis would be easier to use in a dialogic synchronous spoken task. Lexical sophistication is more likely to be predicted by individual proficiency (Kim & Crossley, 2018; Kyle & Crossley, 2015, 2016), as having high WTC does not necessarily allow a learner to access a more complex lexicon.

Overall, it seems that WTC relates to lexical diversity which represents the breadth of language that a learner uses during a task. WTC pushes students to engage more cognitively in the task (i.e., use unique words/avoid repetition) and therefore use the language that they have to communicate effectively. Lexical density and sophistication (representing the depth of language used) may be more constrained by features of the task, and also the language proficiency of learners, and therefore are less suitable as measures of cognitive engagement.

The current study is of interest to both researchers and teachers. First, researchers are increasingly interested in using a variety of different methods to measure student engagement with tasks (Hiver & Wu, 2023). Although behavioral engagement is a useful indicator of task performance, when taken alone it does little to indicate the quality of speech, which is more likely to be related to cognitive engagement. There are also recognized problems with self-reports of task engagement (Hiver & Dao, 2025). The current study showed that cognitive engagement as measured by lexical diversity can be used as an additional objective measure alongside behavioral engagement to give a more rounded view of how students are engaging with tasks. Although the two types of engagement are related, they seem to be measuring different things.

For teachers using a TBLT approach, this study further highlights the importance of WTC as a predictor of language use during tasks. Not only does WTC predict the quantity of language that students will produce, but also appears to be related to the quality of language. This means that teachers need to consider ways in which they can maximize WTC in their students. For example, Yashima et al. (2004) showed that communicative confidence was a positive predictor of WTC. This means that students should be given achievable tasks that will provide them with opportunities for success in the classroom in order to build confidence. Often university contexts attempt to challenge students with complex tasks, but focusing initially on achievable tasks may be more effective.

Conclusion

This exploratory study sought to build on the findings of Leeming et al. (2024) by examining the relationship between WTC and cognitive engagement in a group discussion task. Leeming et al. (2024) measured the total number of words produced (behavioral engagement) and showed that WTC was a significant predictor of this variable. However, behavioral engagement only helps us to understand part of the picture, in that it offers few insights into the quality of the engagement, which is better represented by cognitive engagement. Using the data from Leeming et al. (2024), we were able to show that WTC had a significant relationship with the lexical diversity. This shows that WTC not only predicts the likelihood of learners producing language during tasks, but may also be related to the quality of speech. The addition of lexical complexity increases our understanding of the relationship between WTC, engagement, and language output, and further highlights the importance of WTC as an ID variable in the language classroom. The main limitation of this exploratory finding is that it was the only one of 26 tests to reveal a meaningful association and therefore it could be a "fluke" with weak generalization/external validity. Also, WTC was measured at the trait level and therefore does not account for potential fluctuations in state WTC as influenced by specific task factors.

The current study was only able to establish a correlational relationship, and future research should investigate how self-reported WTC predicts the quality of language produced. This could involve looking not only at the lexical complexity, but also the accuracy and fluency of language. Theoretically, students who are higher in WTC will want to ensure that their language is comprehensible to their interlocutor, and therefore make an effort to be accurate in their use of grammar. This may in turn have a

negative impact on the fluency of language produced (Ellis et al., 2020). Research is needed that investigates these potentially interacting constructs. Finally, there is a need for longitudinal research that tracks development in both WTC and language quality over time. Researchers are aware that even IDs measured at the level of traits (such as in the current study), may still fluctuate over the course of a language course, and longitudinal research designs can capture this dynamic change.

Acknowledgements

This research was support by a grant-in-aid awarded by the Japan Society for the Promotion of Science (21K20214) to the corresponding author. The authors wish to thank Rachel Li and Alessandra Ferrer for their work as research assistants contributing to this project.

References

Abdi, H. (2007). Part (semi partial) and partial regression coefficients. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (Vol. 3, pp. 1–9). SAGE Publications.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, Instruments & Computers, 41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), D*imensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). John Benjamins. http://doi.org/10.1075/lllt.32.02bul

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing, 26*(4), 42–65. https://doi.org/10.1016/j.jslw.2014.09.005

Davies, M. (2008–). *The Corpus of Contemporary American English (COCA): 560 million words, 1990–present*. https://www.english-corpora.org/coca/

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*, 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. Routledge.

Ellis, R., Skehan, P., Li, S., Shintani, N., & Lambert, C. (2020). *Task-based language teaching: Theory and practice*. Cambridge University Press.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/bf03193146

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. MIT Press.

Guiraud, P. (1954). *Les caractéristiques statistiques du vocabulaire* (The statistical characteristics of vocabulary). Presses Universitaires de France.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement* (1st ed.). Routledge. https://doi.org/10.4324/9780203887332

Hiver, P., Al-Hoorie, A. H., Vitta, J. P., & Wu, J. (2024). Engagement in language learning: A systematic review of 20 years of research methods and definitions. *Language Teaching Research*, *28*(1), 201–230. https://doi.org/10.1177/13621688211001289

Hiver, P., & Dao, P. (2025). From task motivation to L2 learning: Understanding links through learners' task engagement. *TESOL Quarterly*. Advance online publication. https://doi.org/10.1002/tesq.3410

Hiver, P., & Wu, J. (2023). Engagement in TBLT. In C. Lambert, S. Aubrey, & G. Bui (Eds.), *The role of the learner in task-based language teaching* (pp. 74–90). Routledge. https://doi.org/10.4324/9781003227267-7

JASP Team. (2025, January 4). *JASP (Version 0.19.3)* [Computer software]. https://jasp-stats.org

Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing, 37*, 39–56. https://doi.org/10.1016/j.asw.2018.03.002

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*, 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12–24. https://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., Crossley, S., & Berger, C. (2018). The Tool for the Automatic Analysis of Lexical Sophistication (TAALES): Version 2.0. *Behavior Research Methods, 50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity using direct judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Leeming, P. (2019). Emergent leadership and group interaction in the task-based language classroom. *TESOL Quarterly*, *53*(3), 768–793. https://doi.org/10.1002/tesq.506

Leeming, P., & Harris, J. (2025). Self-efficacy and collective-efficacy as predictors of engagement in group conversation. *TESOL Quarterly*. Advance online publication. https://doi.org/10.1002/tesq.3412

Leeming, P., Vitta, J. P., Hiver, P., Hicks, D., McLean, S., & Nicklin, C. (2024). Willingness to communicate, speaking self-efficacy, and perceived communicative competence as predictors of second language spoken task production. *Language Learning*, *74*(4), 917–949. https://doi.org/10.1111/lang.12640

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36–62. https://doi.org/10.5054/tq.2011.240859

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232.x

MacIntyre, P. D., Clément, R., Dörnyei, Z., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal*, *82*(4), 545–562. https://doi.org/10.1111/j.1540-4781.1998.tb05543.x

Mercer, S., Talbot, K. R., & Wang, I. K. H. (2021). Fake or real engagement–looks can be deceiving. In P. Hiver, A. H. Al-Hoorie, & S. Mercer (Eds.), *Student engagement in the language classroom* (pp. 143–162). Multilingual Matters.

Pallotti, G. (2009). Defining, refining and differentiating constructs: CAF. *Applied Linguistics, 30*(4), 590–601. https://doi.org/10.1093/applin/amp045

Sato, M., Salas, G., Freeborn, L., & Tajabadi, A. (2025). TBLT, engagement, and group synchronicity: Exploring the relationships between affective, social, and cognitive aspects of task-based performance. *TESOL Quarterly*. Advance online publication. https://doi.org/10.1002/tesq.3408

Teimouri, Y., Plonsky, L., & Tabandeh, F. (2022). L2 grit: Passion and perseverance for second-language learning. *Language Teaching Research*, *26*(5), 893–918. https://doi.org/10.1177/1362168820921895

Vitta, J. P. (2019). *Exploring complexity and text difficulty in a case of CEFR-leveled Cambridge University Press ELT textbooks: Providing guidance for local ELT publishers* (Unpublished doctoral thesis). Queen's University Belfast, School of Social Science, Education, and Social Work.

Vitta, J. P., Leeming, P., & Nicklin, C. (2025). Interpreting the semi-partial correlation as a multiple regression-bound (not a bivariate) metric: A methods-oriented response to Papi and Teimouri's (2024) response to Al-Hoorie et al. (2024). *Studies in Second Language Acquisition*. Advance online publication. https://doi.org/10.1017/S0272263125100946

Vitta, J. P., Nicklin, C., & Albright, S. W. (2023). Academic word difficulty and multidimensional lexical sophistication: An English-for-academic-purposes-focused conceptual replication of Hashimoto and Egbert (2019). *Modern Language Journal, 107*(1), 373–397. https://doi.org/10.1111/modl.12835

Vitta, J. P., Nicklin, C., & McLean, S. (2022). Effect size–driven sample-size planning, randomization, and multisite use in L2 instructed vocabulary acquisition experimental samples. *Studies in Second Language Acquisition, 44*(5), 1424–1448. https://doi.org/10.1017/S0272263121000541

Yashima, T., Zenuk-Nishide, L., & Shimizu, K. (2004). The influence of attitudes and affect on willingness to communicate and second language communication. *Language Learning*, *54*(1), 119–152. https://doi.org/10.1111/j.1467-9922.2004.00250.x