The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance

Andrea Révész

UCL Institute of Education, University College London

a.revesz@ucl.ac.uk

Monika Ekiert

LaGuardia Community College, City University of New York

mekiert@lagcc.cuny.edu

Eivind Nessa Torgersen

Sør-Trøndelag University College

eivind.n.torgersen@hist.no

Abstract

Communicative adequacy is a key construct in second language research, as the primary goal of most language learners is to communicate successfully in real-world situations. Nevertheless, little is known about what linguistic features contribute to communicatively adequate speech. This study fills this gap by investigating the extent to which complexity, accuracy, and fluency (CAF) predict adequacy; and whether proficiency and task type moderate these relationships. Twenty native speakers and 80 second language users from four proficiency levels performed five tasks. Speech samples were rated for adequacy and coded for a range of complexity, accuracy, and fluency indices. Filled pause frequency, a feature of breakdown fluency, emerged as the strongest predictor of adequacy. Predictors with significant but smaller effects included indices of all three CAF dimensions: linguistic complexity (lexical diversity, overall syntactic complexity, syntactic complexity by subordination, frequency of conjoined clauses), accuracy (general accuracy, accuracy of connectors), and fluency (silent pause frequency, speed fluency). For advanced speakers, incidence of false starts also emerged as predicting communicatively adequate speech. Task type did not influence the link between linguistic features and adequacy.

Introduction

Communicative success in the second language (L2) is the primary goal for the majority of L2 learners. For this reason, it appears desirable to define the aim of L2 teaching in terms of preparing learners to be able to communicate adequately in real-world situations aligned with their future academic, professional, and/or personal needs. Based on this rationale, the last two decades have seen a growing body of research investigating various aspects of L2 learners' performance on communicative language tasks. This interest in tasks has been inspired by the fact that pedagogic tasks are meaning-focused and learner-centred, unlike traditional language learning activities which tend to be more decontextualized and grammar-oriented.

A general definition describes tasks as activities "where meaning is primary; there is some communicative problem to solve; some sort of relationship with real-world activities; and the assessment of task is in terms of a task outcome" (Skehan 1998: 95). The construct of communicative task has emerged as a key unit in the areas of L2 teaching and testing. In language teaching, task has been promoted and increasingly used as a curricular unit around which instruction is organised. In many areas of language testing, task is taken as a unit of analysis, which motivates test construction and rating of performances (Brown *et al.* 2002). Motivated by these practical concerns and insights from second language acquisition (SLA) research, tasks and their role in language learning have also become the subject of much theorizing (Skehan 1998; Robinson 2001) and empirical inquiry in instructed SLA.

Despite the importance attributed to tasks as promoters and assessments of communicative adequacy, the bulk of task-related SLA research has been directed at examining the linguistic outcomes of task performance, expressed in terms of syntactic and lexical complexity, accuracy, and fluency, without considering how these features may relate to communicative adequacy (see, however, De Jong *et al*. 2012a, 2012b; Kuiken *et al.* 2010).

The overwhelming focus on learners' lexico-grammar appears a shortcoming (Pallotti 2009), since it is well-known that one can use complex and accurate language while not being functionally effective, and, vice versa, it is possible to get one's message across without using complex language and being accurate. Due to the importance of communicative success in real-world contexts, it appears timely and worthwhile to put more research emphasis on how linguistic factors may facilitate or hinder L2 users' success in completing tasks.

To that end, this study addresses the extent to which linguistic features are linked to communicative adequacy, and whether these relationships differ depending on proficiency level and task type. In particular, the study aims to explore connections between objective measures of speech and ratings of adequacy, the latter understood as the knowledge and employment of both linguistic and interactional resources in social contexts. The novel aspects of our research on adequacy reside in the following: we focused on oral rather than written production (Kuiken *et al.* 2010), employed a wide range of performance measures and considered multiple rather than a single task type.

Communicative Adequacy and Task-Based SLA Research

In the task-based literature, a coherent and clear-cut definition of communicative adequacy as a construct is absent (Kuiken *et al.* 2010). Although adequacy is often used interchangeably with phrases such as "successful performance," "communicative success," "communicative efficacy," or "getting the message through," it is not always clear what individual researchers mean by it. Recently, Pallotti (2009) described adequacy as "the degree to which a learners' performance is more or less successful in achieving the task's goals efficiently" (596). Under Pallotti's definition, adequacy is related to the notion of interactional competence as it involves determining "what a person *does* together with others" (Young 2011: 430). It follows, therefore, that adequacy in the context of spoken interaction refers to the discursive practice, whereby participants recognize and respond to the expectations of what to say and how to say it, contingent on what other participants do and what the context is. We adopted this conceptualisation as a working definition of adequacy for our study.

Another problem in the task-based literature is the fact that few studies report data on whether learners actually succeeded in accomplishing the communicative aims of the task. Thus far, the dominant method has been to measure the success of task-based performance in terms of the learners' use of the language system, involving dimensions of linguistic complexity, accuracy and fluency (CAF). Robinson (2001) refers to this practice as indirect testing. He explains, however, that task performance can also be evaluated in terms of whether the non-linguistic (e.g., a map or list of differences), pragmatic outcome of the task has been accomplished. In a similar vein, De Jong *et al.* (2012a) suggests that assessment exclusively using CAF measures is not sufficient to obtain a valid estimate of successful performance. Ortega (2003) also observes that "progress in a learner's language ability for use may include syntactic complexification, but it also entails the development of discourse and sociolinguistics repertoires that the language user can adapt appropriately to particular communication demands" (493). It is worth noting that the sole use of CAF indices to assess task-based performance is in contrast to the practices of the teaching and testing fields, where the extent to which classroom learners or test-takers have the abilities to function successfully in real-life settings has been given considerable weight.

In the context of task-based research, Pallotti (2009) was one of the first to problematize the exclusive use of CAF measures as benchmarks for successful task performance. He proposed that adequacy should be used, on the one hand, as a separate measure, independent from CAF, and, on the other hand, as a dimension helping to interpret CAF measures. Since Pallotti's seminal paper, the field has witnessed some accumulation of empirical research addressing the issue of how adequacy may relate to CAF indices.

Studies of Linguistic Measures and Communicative Adequacy

The handful of studies that have examined communicative adequacy with respect to linguistic measures have been generated by two projects, the What is Speaking Proficiency (WISP) project and the Communicative Adequacy and Linguistic Complexity (CALC) study. In the WISP project, a large-scale investigation of the componential structure of speaking proficiency, the relationship of communicative or functional adequacy to linguistic knowledge and language skills was explored. The results of two studies emerging from the project are of particular relevance here. De Jong *et al.* (2012b) examined the relative weight of grammatical and vocabulary knowledge, speed of lexical retrieval, articulation, and sentence building, along with pronunciation skills in predicting communicatively adequate L2 performance. All skills, except for articulation indices, were found to be related to adequacy, accounting for 76% of the variation. The researchers, however, identified vocabulary knowledge and correct sentence intonation as the strongest predictors of communicatively adequate speech. Interestingly, the relative contribution of various linguistic skills varied depending on adequacy; a similar increase in linguistic knowledge or processing speed resulted in higher gains for participants rated as more adequate. Using the same dataset, Hulstijn *et al.* (2012) examined the association between communicative adequacy and linguistic competences according to proficiency level. Except for articulation speed, all measures of linguistic knowledge and processing ability were found to discriminate between B1 and B2 levels in terms of the Common European Framework of Reference (CEFR). Importantly, the researchers observed that the differences in lexical and grammatical knowledge were gradual rather than categorical at the two CEFR levels.

Investigations of the extent to which CAF measures predict speaking proficiency also inform our research. In a study of the relationship between holistic ratings of oral proficiency and objective measures of grammatical accuracy and complexity, vocabulary, pronunciation, and fluency, Iwashita *et al.* (2008) found that token frequency (a vocabulary measure) and speech rate (a fluency index) had the strongest impact on speaking proficiency. Additional measures that had a moderate effect on speaking scores included global accuracy (grammatical accuracy), type frequency (vocabulary), target-like syllables (pronunciation), and unfilled pauses and total pause time (fluency). With fluency emerging as a critical component of speaking proficiency, Ginther *et al.* (2010) examined the link between fluency and holistic ratings of speech quality. The study yielded strong and moderate correlations between proficiency scores and indices of speech rate, speech time ratio, mean length of run, and number and length of silent pauses.

To date, the CALC project (Kuiken *et al.* 2010) is the closest to the present study in terms of its aims and design, thus we provide a detailed review of this research. Like the present study, Kuiken *et al.* investigated the link between the linguistic and communicative aspects of L2 performance. Their focus, however, was on written rather than oral production. One hundred and three participants, L2 learners of Dutch, Italian, and Spanish falling within the CEFR A2-B1 proficiency range, completed two open-ended, decision-making tasks. Adequacy was assessed on a six-point scale measuring the writer's ability to fulfil the communicative goal of the task and the impact of the resultant text on the reader. The linguistic complexity of the performances was measured both holistically and with standardized measures of linguistic complexity, accuracy, and fluency. The holistic rating scale comprised seven levels, which were used to rate performances based on general descriptors of syntactic complexity, lexical complexity, and accuracy. Syntactic complexity was expressed as clauses per T-unit and dependent clauses per clause. The accuracy measures included number of errors per 100 words and T-units. Lexical diversity was quantified using Guiraud's Index (a type-token ratio). The results indicated that the correlations between adequacy and linguistic complexity, as measured by rating scales, tended to be higher for more advanced learners. Another finding was that while in most cases the lexical variation and accuracy measures were found to be linked to communicative adequacy, neither of the syntactic complexity indices correlated with adequacy.

Extending Kuiken *et al.*'s work to oral production, the aim of this study was to investigate which linguistic factors facilitate or hinder success in completing oral language tasks in general and at different proficiency levels. Unlike Kuiken *et al.*, we also examined how the association between adequacy and linguistic outcome measures may be influenced by the type of task in which language users engage. The methodological innovation of our research lies in the wide range of linguistic measures employed, including specific measures of linguistic complexity and accuracy.

Research Questions

1. To what extent does linguistic complexity (i.e., syntactic and lexical complexity), accuracy, and fluency predict communicative adequacy during task performance?
2. To what extent does level of proficiency influence the extent to which measures of linguistic complexity, accuracy, and fluency predict communicative adequacy?
3. To what extent does task type influence the extent to which measures of linguistic complexity, accuracy, and fluency predict communicative adequacy?

Methodology

Dataset

The dataset includes performances on five oral tasks by 80 ESL learners and 20 native speakers (NSs) of English, a total of 500 performances. The ESL data were collected as part of a placement test, which was developed and validated at a North-American university for placing students into appropriate levels in the university’s language program (Kim 2006; Purpura 2004) . The test is a theme-based assessment, consisting of five sections: listening, speaking, grammar, reading, and writing. It divides learners into proficiency levels, from beginner to advanced, based on the overall combined score from the five test sections. The speaking score accounts for 25% of the overall score. For this study, the proficiency levels were drawn considering the overall as well as speaking scores. Participants were assigned to a certain level if they met the placement criteria for that level in terms of their overall and speaking score (the correlation between participants' overall and speaking score was very high, *r* = .93). Using these criteria, we selected 20 speakers from four proficiency levels (a total of 80) – low-intermediate (LowInt), intermediate (Int), low-advanced (LowAdv), and advanced (Adv) – from a pool of 600 test-takers. In order to control for L1 differences, 10 Japanese and 10 Spanish learners were randomly chosen per level, given that the majority of the test-takers came from these two L1 backgrounds. The median and mean age of the learners were 29.5 and 31.80 (SD = 7.02), respectively; 75% were female, and 25% were male. Their length of residence in an English speaking country ranged from 11 months to 5 years (M = 2.25, SD = 1.48). One-way ANOVAs run on age and length of residence found no significant differences among test-takers at the four proficiency levels; *F* (3, 76) = .333, *p* = .80; and *F* (3, 76) = .222, *p* = .88, respectively. The median age across the groups was also in a similar range (29.5 - 32). The NSs were specifically recruited for the study, and were all students at the same university. Their average age was 34.55 (SD = 8.23). 70% percent were female, and 30% were male.

Speaking tasks

The five speaking tasks involved making a complaint about a catering service, refusing a suggestion by a teacher, telling a story based on pictures, giving advice based on a radio commentary, and summarizing information from a lecture. They were integrated testing tasks, using various input types. Participants were asked to read, listen, or view the task stimulus, and then to respond to the stimulus when prompted.[1] The tasks were computer-delivered. The planning time varied between 20 to 60 seconds, whereas the speaking times were either 45 or 60 seconds. Prior to testing, participants completed a practice task in which they were asked to introduce themselves. The task order was the same for all participants. The five tasks are summarized in the Supporting Information Online (S1).

Communicative adequacy ratings

The communicative adequacy of the performances was assessed by trained raters. Twenty postgraduate students were recruited, ten doctoral students in linguistics and ten native speakers with no background in linguistics or languages. Our rationale for selecting both linguistically aware and naïve raters was to control for the impact of rater background on rater severity and orientation (e.g., Chalhoub-Deville 1995). Each performance was evaluated by two raters. To ensure sufficient linkage among the ratings (n=1000), we devised the judging plan in such a way that every rater overlapped with every other rater, and each rater assessed performances on each task and from each proficiency level. All raters evaluated 50 performances. The raters completed their ratings in their own time, after they had participated in a training session.

Every sample was rated on a task-independent rating scale, which was accompanied by task-dependent content points. The task-independent scale consisted of seven levels and included descriptors related to whether the speaker addressed and supported by sufficient detail the task-specific content points, was easy or difficult to understand, delivered the message in a clear and effective manner, and took account of the communicative situation (see Supporting Information Online S2). The task-dependent content points described elements which were essential to task completion. For example, the content points for the task which asked participants to refuse a teacher's suggestion were as follows: (a) acknowledge receipt of phone message and/or professor's opinion, (b) express disagreement with professor's position, and (c) make case for own position and/or solution. The development of the rating instruments was informed by previous research (De Jong *et al*. 2012a; Brown *et al.* 2002; Tankó 2005) and the expert opinions of language teachers and testing experts, the majority of whom were associated with the program where the data were collected.

Linguistic analyses of speaking performances

The 500 performances were transcribed by one of the researchers using PRAAT (Boersma and Weenink 2007). Ten percent of the transcripts were checked by another researcher, yielding an inter-transcriber agreement of 98%. Next, the samples were analyzed in terms of linguistic complexity (i.e., syntactic and lexical complexity), accuracy, and fluency.

Lexical complexity was assessed using measures of lexical frequency, lexical density, and lexical diversity. Lexical frequency was gauged by the means of Web VocabProfile v3 (Cobb n.d.). This program calculated the percentage of words, function words, and content words that were among the 1000 most frequent English word families (K1 words, function words, and content words), the percentage of words contained in the 1000-2000 most frequent word families (K2 words), and the percentage of words included in the 2000 frequency band (K1 + K2 words). The program also computed the percentage of words belonging to *The Academic Word List* (Coxhead 2000), and the percentage of words that did not appear in any of these lists. Lexical density was also obtained using the program Web VocabProfile v3 (Cobb n.d.), and was expressed as the proportion of content words to the total number of words. We measured lexical diversity, the range and variety of words in a text, by Malvern and Richards' (1997) D-formula, a type-token ratio that statistically controls for text length. Given that the program can only be used for texts longer than 50 tokens, we were unable to obtain D for 8.8% of our dataset (n = 44 samples) since the performances were of shorter length. It was also not possible to supplement D with the measure of textual lexical diversity (MTLD) as recommended by McCarthy and Jarvis (2010). The MTLD tool requires a minimum of 100 tokens, and 43.7% of our dataset had fewer tokens.

To take account of the multi-faceted nature of syntactic complexity, the speaking performances were evaluated in terms of general and specific syntactic complexity measures. We used three types of general indices: subordination, phrasal, and overall complexity (Norris and Ortega 2009). Complexity by subordination was expressed as the proportion of clauses to analysis of speech units (AS-units, Foster *et al.* 2000). To assess phrasal complexity, the number of words in each sample was divided by the number of clauses in the sample. As a measure of overall complexity, the ratio of words to AS-units was calculated. In coding for specific measures, we obtained the frequency (number of tokens per 100 words) and Guiraud's index (GI = type/squareroot of token) for tense-aspect forms, modal verbs, and type of clauses (additive, temporal, causal, logical, relative).

Like syntactic complexity, accuracy was assessed based on general and specific measures. As a general index, the proportion of errors per 100 words was calculated. We coded for errors in grammar (e.g., I *am not* agree with you.) and lexis (e.g., pass the course with a great *note*). To gauge the performances in terms of specific features, we examined the extent to which participants used subject-verb agreement, tense-aspect forms, modal verbs, and connectors correctly. For specific grammatical features, scores of suppliance in obligatory contexts (SOC, Brown 1973) were obtained to account for under-suppliance. Except for subject-verb agreement, scores of target-like use (TLU, Pica 1983) were also computed to capture instances of over-suppliance.

To assess various aspects of fluency, we obtained indices of breakdown, speed, as well as repair fluency (Skehan 2003). Breakdown fluency, which measures silence and pausing behaviour, was assessed by silent pauses and filled pauses (uhms and uhs) per 100 words. Pauses were defined as silent periods exceeding 250 ms, a cut-off point often employed to distinguish pauses from hesitation (Goldman-Eisler 1968). Speed fluency is a measure of the speed of one's speech. Following De Jong *et al.* (2013), speed fluency was operationalized as inverse articulation rate or mean duration of syllables, which we obtained by dividing speaking time (excluding pauses) with the number of syllables. Repair fluency, the frequency with which speakers use false starts, and repeat and repair their utterances, was expressed in terms of false starts per 100 words, self-repairs per 100 words, and repetitions per 100 words.

The samples were coded by trained research assistants. To check reliability, four participants (20% of the data) were randomly selected from each of the five levels (four L2 levels and NSs) and their speech samples were coded by one of the researchers. Inter-coder reliability was high for all coding categories ($.92 < r < .97$, $p < .01$).

Statistical analyses

To estimate the communicative adequacy of the 500 speaking performances as assessed by the 20 raters, the simple Rasch (1960) model was applied using the program FACETS. This analysis converted the raw ratings into their natural logarithm or log-odds (logits), and produced measures for the two facets of the model – communicative adequacy and rater severity – on a true interval scale, known as the *logit scale*. Our rationale for using this model was that it controlled for differences in rater severity when calculating the adequacy of the speech samples, thus resulting in more reliable adequacy estimates. The Rasch measurement also computed *fit statistics* for each element of the two facets, which indicated how well the data fit the stochastic expectation of the model. These fit statistics, for instance, were used to examine how consistently a particular rater assessed communicative adequacy. The Rating Scale model was used for the analysis, which assumes that the steps of a scale are equivalent across all elements of a given facet.

To address the research questions, a series of linear mixed effects regression analyses were conducted, using the lme4 package within the R statistical programming environment. Given that each participant carried out five tasks, multilevel mixed modelling was performed where the variable *task* was nested within the variable *participant*. Hence, the effect of clustering of one variable within another was accounted for in the resulting two-level models. Task and participant served as random effects in each of the analyses (with task nested within participant). The fixed effects in the models varied according to the research question addressed. Using Bonferroni's adjustment, an alpha level of $p < .002$ was set for all tests in order to decrease the possibility of a Type 1 error (.002 = .05 / 32 predictor variables). To measure effect sizes, we obtained marginal R^2 values (variance explained by fixed effects only) and conditional R^2 values (variance explained by fixed and random effects) using the R package MuMIn. Standard diagnostic procedures were used to ensure the appropriateness of the Rasch and regression analyses. In cases where the distributions for the predictor variables were found to be skewed, the analyses were also run with the data transformed into logarithmic values. Given that the analyses including the transformed and raw data did not yield different trends, the results with the raw data are reported here to ease interpretation.

Results

Communicative adequacy and rater severity: Results from Rasch analysis

First, descriptive statistics were computed based on the raw adequacy ratings. The mean adequacy score was 4.61 (SD = 2.08), indicating considerable variation among the adequacy of the speech samples. As expected, participants with higher proficiency achieved higher adequacy scores (low-intermediate: M = 2.58, SD = 1.59, intermediate: M = 3.34, SD = 1.81, low-advanced: M = 4.93, SD = 1.61, advanced: M = 5.69, SD = 1.38), with the native speakers' performances being rated as most adequate (M = 6.46, SD = 1.03). There was smaller variation in adequacy across the five tasks (complaint: M = 4.42, SD = 2.10, refusal:

M = 4.85, SD = 2.05, narrative: M = 4.74, SD = 2.01; advice: M = 4.95, SD = 1.91, summary: M = 4.13, SD = 2.24).

The results of the Rasch analysis (see the Rasch map in S3) confirmed large variation in the communicative adequacy of the samples. The adequacy estimates ranged from -7.80 to 8.05 logits, with a mean and standard deviation of 1.87 and 3.24, respectively. The overall difference between the adequacy estimates was significant, $\chi^2(499) = 3135.3$, $p < .01$. The separation reliability, which corresponds to Cronbach's alpha, was .88, indicating that participants can be separated into different categories with good reliability. These indices suggest that the adequacy of the performances was spread out on the logit scale consistently. The infit statistics, which identify irregular ratings (e.g., a sample being rated as more adequate by a severe rater than a lenient rater), show that the majority (94%) of the ratings had infit values in the acceptable range of two standard deviations (SD = 1.14) around the mean (M = .86) (Pollitt and Hutchinson 1987).

For rater severity, the mean was set at 0 logits, and the analysis yielded a standard deviation of 1.15 logits. The raters ranged in severity from -.79 to 2.35 logits. The overall difference among raters was significant, $\chi^2(18) = 552.0$, $p < .01$, with a separation reliability of .97. The infit mean square values were all in the acceptable range of .50 to 1.50 (Linacre 2002) after one misfitting rater had been removed after preliminary analyses. These results indicate that the self-consistency of the raters was acceptable in assessing communicative adequacy. Of note, raters with a background in linguistics were slightly more severe (M = .29) than raters who were linguistically naïve (M = -.26). The infit mean squares for both groups were close to the Rasch-modeled expectation of 1 (linguists: M = 1.12, SD = .33; non-linguists: M = .92, SD = .37), indicating that, overall, there was little difference in the self-consistency of linguist and non-linguist raters.

CAF and communicative adequacy: Results from linear mixed effects regression analyses

In examining linguistic complexity, accuracy, and fluency of the performances, the data were first checked for outliers for all measures. Outliers were defined as values more than three standard deviations away from the mean. Tables 1-4 provide the descriptive statistics for lexical complexity, syntactic complexity, accuracy , and fluency after outliers were removed. Each table presents the results by proficiency, task type, and total score reflecting the research questions.

TABLES 1-4 ABOUT HERE

The first research question asked whether linguistic complexity, accuracy, and fluency predicted communicative adequacy. We addressed this question by conducting a series of multilevel linear mixed effects regression analyses. The adequacy estimates from the Rasch analysis served as the dependent variable in each analysis. The fixed effect was one of the linguistic complexity, accuracy, or fluency measures. The variables *task* and *participant* were set as random effects, with *task* nested within *participant*. Table 5 presents the statistics for the analyses in which the fixed effect, our predictor of interest, emerged as significant. As shown in Table 5, breakdown fluency, as assessed by incidence of filled pauses, was found to be the strongest predictor of communicative adequacy. As an individual factor, this variable explained 15% of the variability ($R^2 = .15$). Speaking performances were rated as more adequate if they contained fewer filled pauses. The rest of the significant predictors had a considerably smaller influence on communicative adequacy, accounting for not more than 7% of the variability as individual factors ($.01 < R^2 < .07$). These predictors with small effects included indices of all CAF dimensions: linguistic complexity (lexical diversity, overall syntactic complexity, subordination complexity, frequency of coinjoined clauses), accuracy (general accuracy, accuracy of connectors), and fluency (breakdown and speed fluency). Performances received higher communicative adequacy ratings if they were more lexically diverse, syntactically complex, accurate, and fluent.

INSERT TABLE 5 ABOUT HERE

As a follow-up analysis, we ran an additional multilevel linear mixed effects regression analysis, which included the factors that were found to be significant individual predictors of adequacy. These CAF indices were modelled as fixed effects. Similar to the previous regression analyses, the dependent variable was communicative adequacy in the model, and task and participant served as random effects (task nested within participant). The CAF measures (fixed effects) accounted for 41% of the variation among the adequacy estimates (R^2 = .41, p < .001), whereas the overall model (including fixed and random effects) explained 57% of the variance (R^2 = .57, p < .001).

The second research question focused on whether proficiency moderated the relationships between adequacy and the CAF measures – that is, whether the relationship between communicative adequacy and the CAF measures differed depending on proficiency. To address this question, a series of multilevel linear mixed effects regression analyses were performed. First, the Rasch adequacy estimates were set as dependent variables, with one of the CAF measures, proficiency level, and their interaction serving as the fixed effects. Task and participant were modelled as random effects, where task was nested within participant. Proficiency was added to the model as an ordinal variable (LowInt = 1; Int = 2; LowAdv = 3; Adv = 4; Native = 5), reflecting the scalar nature of the construct. In the 32 multiple regression analyses performed, the predictor of interest was the interaction effect between the

CAF measure and proficiency. If a significant interaction was found, this would mean that proficiency level influenced the extent to which a CAF measure predicted adequacy.

A significant interaction effect emerged from only one regression, containing the repair fluency measure of incidence of false starts (*Est* = -13.95, *SE* = 3.99, *t* = 3.49, *p* < .001). Thus, the impact of repair fluency on adequacy was significantly different depending on proficiency. To investigate the interaction effects further, simple multilevel mixed effects regression analyses were carried out separately for the five proficiency levels. In each analysis, communicative adequacy was used as the dependent variable, incidence of false starts served as the fixed effect, and task and participants were set as random effects. Incidence of false starts emerged as a significant, positive predictor only for advanced speakers (fixed effect: *Est* = -13.95, *SE* = 3.99, *t* = 3.49, *p* < .001; random effects: *Var* (Participant) = .97, *SD* (Task) = .98; *Var* (Task) = .09, *SD* (Task) = .31). It proved to be a moderate predictor, with an individual contribution of 15% to the variance in communicative adequacy (R^2 = .15). In sum, a lower incidence of false starts was associated with higher adequacy for advanced speakers.

The third research question asked whether relationships between communicative adequacy and the CAF measures were moderated by task – that is, whether the CAF measures differentially predicted adequacy depending on task. The same statistical procedures were followed as in investigating the second research question. First, a series of multiple multilevel linear mixed effects regressions were performed, with the Rasch adequacy ratings serving as the dependent variable and the CAF measure, task type, and their interaction set as the fixed effects. Our predictor of interest was the interaction between task type and the relevant CAF measure. The random effects were task and participant, with task nested within participant. None of the 32 multiple linear mixed effects regressions yielded a significant interaction.

Discussion and Conclusion

This study investigated which linguistic factors may facilitate or hinder success in completing oral language tasks, recognizing the important role that tasks may play in promoting and assessing communicative adequacy. The study also examined how the potential association between adequacy and linguistic outcome measures may be influenced by proficiency and the task in which language users engage.

We found that a set of linguistic factors had a significant impact on communicative adequacy as perceived by trained raters. Frequency of filled pauses, a feature of breakdown fluency, emerged as the strongest predictor. Eight additional features were found to have significant but weaker relationships with adequacy, including lexical diversity, overall syntactic complexity, subordination complexity, conjoined clause frequency, general accuracy, connector accuracy, silent pause frequency, and speed fluency. In other words, fluency emerged as a critical determinant of communicative adequacy. The fact that filled pause frequency, a breakdown fluency measure, had the strongest effect on adequacy is in line with the findings of fluency research, where perceived fluency judged by raters is consistently found to have strong associations with breakdown fluency (see Bosker *et al.* 2013). In future studies, it would be interesting to explore the extent to which ratings of adequacy and fluency may be related, given Freed’s (1995) suggestion that, when assessing fluency, raters probably take other performance aspects into account, in addition to actual fluency indicators.

Interestingly, repair fluency was the only CAF measure that showed differential impact on communicative adequacy depending on proficiency. Higher adequacy was found to be associated with lower incidence of false starts in the advanced L2 users’ speech. False starts occurred rarely and with about the same frequency across proficiency levels, but, since advanced users demonstrated more superior skills in other linguistic areas, the presence of false starts may have become more noticeable and distracting in their performance.

As regards linguistic complexity and accuracy, our findings replicate those of Kuiken *et al.* (2010) for lexical diversity and general accuracy, but we obtained different results for syntactic complexity. Like Kuiken *et al.*, we found that the speakers' ability to achieve the task goals efficiently was associated with the use of more diverse lexis and accurate language. However, contrary to Kuiken *et al*., we also identified subordination complexity as a significant predictor of adequacy. The discrepancy between Kuiken *et al.*'s and our findings may lie in our use of the oral rather than the written mode. When processing written language, grammatical errors and limited lexis are probably perceived as more disruptive than simple syntax. For example, failure to supply certain grammatical markers in writing is more likely to capture raters' attention than in speaking where, due to the phonetic realisations of relevant forms, grammatical errors become less salient. The positive effects of subordination on adequacy in speech may be accounted for by the fact that, in the context of oral language, where subordination is less frequent than in writing, the facilitative effect of subordination on logical and temporal cohesion becomes more salient, having a positive impact on the perception of communicatively adequate speech by raters. Our finding that, in addition to general accuracy, the accuracy of connectors played a significant role in predicting adequacy seems to support this line of reasoning.

We also looked into the extent to which the variance in the communicative adequacy ratings can be explained by all the significant predictors together. The multiple mixed effects regression model we ran including all significant factors accounted for 57% of the variance in adequacy (fixed and random effects), of which 41% could be attributed to CAF measures (fixed effects). It is interesting to compare this finding to that of De Jong *et al.* (2012b), who were able to explain 76% of variation in communicative adequacy, using grammatical and vocabulary knowledge, speed of lexical retrieval, articulation, sentence building, and pronunciation skills as predictors. A possible explanation why our model was able to explain less variance may lie in the fact that we did not consider pronunciation quality, while this factor had a strong impact on adequacy in De Jong *et al.* (2012b). Another important difference between our and the DeJong *et al*'s research is that we considered CAF performance measures, whereas DeJong *et al* investigated the contribution of underlying linguistic knowledge and processing skills to adequacy.

Task type was not found to moderate the relationship between adequacy and the CAF measures. This finding, however, needs to be treated with caution, given that we only considered a limited number of specific constructions. It would be worthwhile to explore additional relationships between specific linguistic features and communicative adequacy, given Loschky and Bley-Vroman’s (1993) proposal that, as a function of task design, particular constructions may be of more utility for achieving successful performance. For example, a successful response on the narrative might be associated with the successful use of temporal connectives and past tense forms, specific linguistic features we did not code for.

Finally, it is also worth considering our findings in relation to previous studies investigating the links between objective measures of CAF and overall speaking proficiency test scores which are often defined in terms of language—in addition to adequacy—related descriptors. In our dataset, communicative adequacy and overall speaking proficiency were closely linked; a follow-up Spearman correlation computed between the adequacy estimates and the overall speaking proficiency scores yielded a strong positive relationship (ρ = .66). In light of this, it is not surprising that a number of CAF indices found to have a significant impact on communicative adequacy here also emerged as moderate to strong predictors of speaking proficiency scores in Iwashita *et al.* (2008). These indices include breakdown and speed fluency, lexical diversity, and general accuracy. Our results are also well aligned with those of Ginther *et al.* (2010), who identified significant relationships between speaking proficiency and measures of speed and breakdown fluency.

There are a number of limitations to this study that need to be acknowledged and addressed in further research. First, given the large number of predictor variables, our dataset was not sufficiently large to be analyzed using more sophisticated statistical procedures such as structural equation modelling. Second, as mentioned above, we only looked into the relationship between a limited number of specific linguistic constructions and communicative adequacy. In future research, it would be interesting to explore this relationship by selecting linguistic features that are relevant to successful task completion (e.g., temporal connectives in narratives). Third, our dataset would have lent itself well to investigating the validity of the CAF measures used, given that the participant pool included both native speakers and L2 users from various proficiency levels. This was beyond the scope of this study, but is a worthwhile future research direction. Despite these limitations, the findings of our study have yielded valuable new insights, and confirmed that the exploration of communicative adequacy in relation to linguistic measures is an important research endeavour.

Endnotes

[1] An anonymous reviewer pointed out that participants' listening ability might have moderated the results, given that some of the tasks required processing aural input. A Pearson correlation computed between participants' listening and speaking placement test scores revealed a strong correlation ($r = .77$, $n = 80$, $p < .01$), suggesting that differences in listening ability were unlikely to have a considerable impact on the findings.

Acknowledgments

The authors would like to thank the Community Language Program at Teachers College, Columbia University and Drs James Purpura and K. Philip Choong for making participant data available for use in this research. We would also like to thank the following individuals for their assistance in coding the data: Chihiro Inoue, Janina Iwaniec, Sofia Lampropoulou, Diana Mazgutova, Sharon McCulloch, Charis Stefanou, and Uschie Weinberger. This research was supported in part by Lancaster University's Early Career Research Grant and PSC-CUNY Research Award No. 65331-00-43. Finally, we are grateful to the anonymous reviewers and the editors of *Applied Linguistics* for their helpful suggestions on the manuscript.

References

Boersma, P. and **D. Weenink**. 2007. Praat: Doing Phonetics by Computer. Software version 4.6.09. www.praat.org.

Bosker, H. R., Pinget, A. F., Quené, H., Sanders, T. and **de Jong, N. H.** 2013. 'What makes speech sound fluent? The contributions of pauses, speed and repairs.' *Language Testing* 30/2: 159-175.

Brown, R. 1973. *A First Language: The Early Stages*. London: George Allen & Unwin.

Brown, J. D., **T. Hudson**, **J. M. Norris** and **W. Bonk** 2002. *An Investigation of Second Language Task-based Performance Assessments.* (Technical Report #24). Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center.

Chalhoub-Deville, M. 1995. 'Deriving oral assessment scales across different tests and rater groups.' *Language Testing* 12/1: 16-33.

Cobb, T. (n.d.). Web Vocabprofile. Retrieved from http://www.lextutor.ca/vp/eng, an adaptation of Heatley & Nation's (1994) Range .

Coxhead, A. 2000. ‘A new academic word list.’ *TESOL Quarterly* 34/2: 213-238.

De Jong, N., M. Steinel, A. Florijn, R. Schoonen and **J. Hulstijn.** 2012a. ‘The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and nonnative speakers’ in A. Housen, F. Kuiken and I. Vedder (eds.): *Dimensions of L2 Performance and Proficiency. Investigating Complexity, Accuracy and Fluency in SLA.* Amsterdam: John Benjamins.

De Jong, N., M. Steinel, A. Florijn, R. Schoonen and **J. Hulstijn.** 2012b. ‘Facets of speaking proficiency.’ *Studies in Second Language Acquisition* 34/1: 5-34.

De Jong, N., M. P. Steinel, A. F. Florijn, R. Schoonen and **J. Hulstijn.** 2013. ‘Linguistic skills and speaking fluency in a second language.’ *Applied Psycholinguistics 34*: 893-916.

Foster, P., A. Tonkyn and **G. Wigglesworth.** 2000. ‘Measuring spoken language: A unit for all reasons.’ *Applied Linguistics* 21/3: 354-375.

Freed, B. 1995. ‘Do students who study abroad become fluent?’ in B. Freed (ed.): *Second Language Acquisition in a Study Abroad Context.* Amsterdam: John Benjamins.

Ginther, A., S. Dimova, and **R. Yang.** 2010. ‘Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring.’ *Language Testing* 27/3: 379-399.

Goldman-Eisler, F. 1968. *Psycholinguistics: Experiments in Spontaneous Speech*. New York: Academic Press.

Hulstijn, J., R. Schoonen, N. de Jong, M. Steinel and **A. Florijn**. 2012. ‘Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR).’ *Language Testing* 29/2: 202-220.

Iwashita, N., A. Brown, T. McNamara and **S. O'Hagan.** 'Assessed levels of second language speaking proficiency: How distinct?' *Applied Linguistics* 29/1: 24-49.

Kim, H-J. 2006. 'Providing validity evidence for a speaking test using FACETS.' *Teachers College Columbia University Working Papers in TESOL & Applied Linguistics* 6/1: http://journals.tc-library.org/index.php/tesol/article/view/180

Kuiken, F., I. Vedder and **R. Gilabert.** 2010. 'Communicative adequacy and linguistic complexity in L2 writing,' in I. Bartning, M. Martin and I. Vedder (eds): *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research***.** Eurosla Monographs 1. Roma: Eurosla.

Linacre, M. 2002. 'Optimizing rating scale category effectiveness.' *Journal of Applied Measurement* 3/1: 85-106.

Loschky, L. and **R. Bley-Vroman** 1993. Grammar and task-based methodology, in G. Crookes and S. Gass (eds.): *Tasks and Language Learning: Integrating Theory and Practice*. Clevedon, England: Multilingual Matters Ltd.

Malvern, D. and **B. Richards**. 1997. 'A new measure of lexical diversity,' in A. Ryan and A. Wray (eds): *Evolving Models of Language*. Clevedon: Multilingual Matters.

McCarthy, P. and **S. Jarvis.** 2010. 'MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment.' *Behavior Research Methods* 42/2: 381-392.

Norris, J. and **L. Ortega.** 2009. 'Towards an organic approach to investigating CAF in instructed SLA: The case of complexity,' *Applied Linguistics* 30/4: 555-578.

Ortega, L. 2003. 'Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing.' *Applied Linguistics* 24: 492-518

Pallotti, G. 2009. 'CAF: Defining, refining and differentiating constructs.' *Applied Linguistics* 30/4: 590-601.

Pica, T. 1983.'Methods of morpheme quantification: Their effect on the interpretation of second language data,' *Studies in Second Language Acquisition* 6: 69-78.

Politt, A., and **C. Hutchinson**. 1987.' Calibrated graded assessment: Rasch partial credit analysis of performance in writing.' *Language Testing* 4/1: 72-92.

Purpura, J. 2004. *Assessing Grammar.* Cambridge: Cambridge University Press.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment tests.* Copenhagen: Danish Institute of Educational Research.

Robinson, P. 2001. 'Task complexity, task difficulty, and task production: Exploring interactions in a componential framework.' *Applied Linguistics* 22/1: 27-57.

Skehan, P. 1998. *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Skehan, P. 2003. 'Task-based instruction,' *Language Teaching* 36/1: 1-14.

Tankó, Gy. 2005. *Into Europe: The Writing Handbook.* Budapest: Teleki László Foundation.

Young, R. 2011. 'Interactional competence in language learning, teaching, and testing,' in E. Hinkel (ed.): *Handbook of Research in Second Language Teaching and Learning.* London: Routledge.

Table 1
Lexical Complexity Results by Proficiency, Task, and Total

Construct	Measure	N	Proficiency Level					Task Type					Total
			LowInt	Int	LowAdv	Adv	Native	Comp	Ref	Nar	Adv	Sum	M
			M	M	M	M	M	M	M	M	M	M	SD
			SD	SD	SD	SD	SD	SD	SD	SD	SD	SD	
Lexical range	K1 words	497	.88	.88	.88	.87	.88	.89	.91	.86	.87	.86	.88
			.05	.05	.04	.04	.03	.04	.04	.04	.04	.05	.04
	K1 function words	498	.58	.58	.57	.57	.57	.58	.60	.59	.56	.54	.57
			.08	.07	.05	.06	.04	.06	.06	.05	.05	.05	.06
	K1 content words	496	.29	.30	.31	.29	.31	.30	.31	.27	.31	.32	.30
			.06	.06	.05	.05	.02	.05	.05	.05	.05	.05	.05
	K2 words	496	.05	.05	.05	.05	.04	.05	.04	.07	.04	.03	.05
			.03	.03	.03	.03	.02	.03	.03	.03	.02	.02	.03
	K1+K2 words	495	.93	.93	.93	.92	.92	.94	.95	.94	.91	.89	.93
			.05	.04	.04	.04	.04	.03	.03	.03	.04	.05	.04
	Academic words	487	.01	.01	.01	.02	.02	.01	.02	.01	.01	.02	.01
			.01	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01
	Off-list words	494	.06	.06	.06	.06	.06	.05	.03	.05	.08	.09	.06
			.04	.04	.04	.04	.03	.03	.03	.03	.03	.04	.04
Lexical density	Cont. words/total words	498	.42	.42	.43	.43	.43	.42	.40	.41	.44	.46	.43
			.07	.07	.05	.06	.04	.06	.06	.05	.06	.05	.06
Lexical diversity	D-value	450	29.61	37.49	42.67	43.66	62.46	5.45	38.68	38.96	48.93	43.38	44.00
			9.15	13.14	12.07	14.06	17.48	2.85	13.65	13.12	17.58	17.19	17.27

Note: Comp = Complaint, Ref = Refusal, Nar = Narrative, Sum = Summary

Table 2
Syntactic Complexity Results by Proficiency, Task, and Total

			Proficiency Level					Task Type				Total
Construct	Measure	N	LowInt	Int	LowAdv	Adv	Native	Comp	Ref	Nar	Adv	Sum	M
			M	M	M	M	M	M	M	M	M	M	SD
			SD	SD	SD	SD	SD	SD	SD	SD	SD	SD	
Subordiation	Clause/AS-unit	492	2.17	2.18	2.07	2.26	3.06	2.13	2.65	2.48	2.48	2.01	2.35
			.87	.75	.54	.52	.79	.69	.90	.75	.68	.74	.79
Phrasal	Words/clause	494	5.13	5.50	5.76	5.95	5.83	5.46	5.05	5.68	5.72	6.31	5.64
			1.04	.99	.93	.95	.78	.89	.83	.78	.94	1.04	.98
Overall	Words/AS-unit	498	11.64	11.75	11.98	13.34	17.85	11.37	14.00	13.99	14.38	12.78	13.31
			4.98	3.63	3.27	3.06	4.40	3.50	5.41	4.27	4.40	4.45	4.57
Frequency: spec. forms	Tense-aspect forms /100 words	495	.14	.14	.13	.12	.12	.13	.12	.16	.11	.13	.13
			.05	.04	.04	.04	.02	.04	.04	.03	.03	.04	.04
	Modal verbs/100 words	498	.02	.02	.02	.02	.02	.02	.03	.00	.03	.01	.02
			.02	.02	.02	.02	.02	.02	.02	.01	.02	.01	.02
	Conjoined clauses /100 words	497	.15	.17	.17	.16	.17	.15	.16	.21	.17	.13	.16
			.08	.06	.07	.06	.04	.06	.05	.06	.06	.06	.06
Diversity: spec. forms	GI for tense-aspect forms	494	.75	.83	.80	.84	.86	.98	.87	.72	.75	.76	.82
			.31	.29	.29	.26	.25	.26	.32	.22	.26	.27	.28
	GI for modal verbs	498	.55	.61	.75	.79	1.07	.90	1.07	.36	1.10	.34	.76
			.59	.62	.62	.68	.65	.59	.67	.57	.49	.48	.66
	GI for conjoined clauses	500	.73	.80	.79	.74	.80	.78	.78	.78	.78	.74	.77
			.28	.22	.17	.18	.16	.17	.21	.18	.20	.27	.21

Note: Comp = Complaint, Ref = Refusal, Nar = Narrative, Sum = Summary

Table 3
Accuracy Results by Proficiency, Task, and Total

			Proficiency Level					Task Type					Total
Construct	Measure	N	LowInt	Int	LowAdv	Adv	Native	Comp	Ref	Nar	Adv	Sum	M
			M	M	M	M	M	M	M	M	M	M	SD
			SD	SD	SD	SD	SD	SD	SD	SD	SD	SD	
General	Errors/100 words	500	.11	.07	.08	.08	.01	.06	.06	.07	.08	.08	.07
accuracy			.05	.04	.04	.03	.01	.04	.04	.05	.05	.05	.05
Underuse:	SOC: Subject-verb	492	.94	.95	.95	.95	.99	.98	.98	.91	.96	.95	.96
spec. forms	agreement		.09	.08	.08	.08	.02	.06	.06	.10	.06	.08	.08
	SOC: Tense-aspect	490	.88	.84	.86	.86	.99	.85	.93	.81	.97	.87	.89
			.16	.16	.13	.13	.04	.15	.10	.17	.06	.15	.14
	SOC: Modal verbs	500	.50	.51	.62	.61	.79	.73	.79	.28	.89	.35	.61
			.48	.49	.47	.47	.40	.43	.39	.44	.27	.48	.47
	SOC: Connectors	489	.92	.92	.92	.93	1.00	.94	.95	.94	.95	.92	.94
			.15	.13	.12	.10	.02	.12	.11	.09	.11	.14	.12
Overuse:	TLU: Tense-aspect	496	.80	.77	.77	.79	.97	.77	.90	.70	.94	.79	.82
spec. forms			.25	.22	.21	.19	.06	.20	.14	.24	.09	.23	.21
	TLU: Modal verbs	500	.49	.51	.62	.61	.78	.72	.78	.28	.89	.35	.60
			.48	.49	.47	.47	.41	.43	.39	.44	.27	.48	.47
	TLU: Connectors	489	.91	.92	.92	.93	1.00	.94	.94	.94	.94	.92	.94
			.15	.13	.12	.11	.02	.13	.11	.09	.11	.14	.12

Note: Comp = Complaint, Ref = Refusal, Nar = Narrative, Sum = Summary

Table 4
Fluency Results by Proficiency, Task, and Total

					Proficiency Level					Task Type			Total
Construct	Measure	N	LowInt	Int	LowAdv	Adv	Native	Comp	Ref	Nar	Adv	Sum	M
			M	M	M	M	M	M	M	M	M	M	SD
			SD	SD	SD	SD	SD	SD	SD	SD	SD	SD	
Breakdown	Silent pauses/100	494	.17	.21	.20	.16	.03	.15	.13	.16	.14	.19	.15
fluency	words		.018	.16	.12	.10	.02	.14	.12	.14	.12	.18	.14
	Filled pauses/100	491	.15	.12	.08	.08	.04	.09	.09	.08	.08	.12	.09
	words		.10	.10	.07	.06	.04	.09	.09	.08	.08	.09	.09
Speed	Speaking time/	494	.26	.23	.23	.23	.22	.23	.23	.23	.23	.24	.23
fluency	syllables		.04	.03	.03	.03	.02	.03	.03	.03	.03	.03	.03
Repair	False starts/100	491	.02	.02	.02	.02	.01	.02	.02	.02	.02	.02	.02
Fluency	words		.02	.02	.02	.02	.01	.02	.02	.02	.02	.02	.02
	Self-repairs/100	494	.02	.02	.02	.02	.00	.02	.01	.02	.01	.02	.02
	words		.02	.02	.02	.02	.01	.02	.02	.02	.02	.02	.02
	Repetitions/100	488	.06	.05	.04	.04	.02	.04	.05	.04	.04	.04	.04
	words		.05	.05	.04	.04	.02	.04	.04	.05	.04	.04	.04

Note: Comp = Complaint, Ref = Refusal, Nar = Narrative, Sum = Summary

Table 5

Estimated Coefficients from Simple Multilevel Mixed Effects Models for Linguistic Complexity, Accuracy, and Fluency Predicting Communicative Adequacy

	Fixed effect statistics					Random effects statistics			
Fixed effect in model	Estimate	SE	T	p	R^2	participant SD	task SD	residual SD	Total R^2
Lexical complexity									
D-value	.04	.01	4.71	<.001	.05	1.87	.02	1.96	.52
Syntactic complexity									
Clause/AS-unit	.83	.15	5.40	<.001	.04	2.07	.08	1.99	.60
Words/AS-unit	.16	.03	6.03	<.001	.06	1.83	.13	1.98	.58
Conjoined clauses/100 words	6.42	1.61	4.01	<.001	.02	2.22	.14	1.96	.63
Accuracy									
Errors/100 words	-15.50	2.68	-5.79	<.001	.06	1.90	.09	2.00	.57
SOC: Connectors	3.37	.90	3.75	<.001	.02	2.08	.10	1.95	.61
TLU: Connectors	3.04	.89	3.40	<.001	.01	2.10	.10	1.96	.61
Fluency									
Silent pauses/100 words	-4.16	1.00	-4.15	<.001	.04	2.10	.09	1.98	.61
Filled pauses/100 words	-13.81	1.66	-8.30	<.001	.15	1.71	.12	1.94	.60
Speaking time/syllables	-26.12	4.39	-5.95	<.001	.07	1.84	.13	1.99	.59