JALT Journal

JALT Journal is the research journal of the Japan Association for Language Teaching (JALT). It is published semiannually, in May and November. As a nonprofit organization dedicated to promoting excellence in language learning, teaching, and research, JALT has a rich tradition of publishing relevant material in its many publications.

Links

- JALT Publications: http://jalt-publications.org
- *JALT Journal*: http://jalt-publications.org/jj
- *The Language Teacher*: http://jalt-publications.org/tlt
- *Conference Proceedings*: http://jalt-publications.org/proceedings

- JALT National: http://jalt.org
- Membership: http://jalt.org/main/membership

JALT
THE JAPAN ASSOCIATION FOR LANGUAGE TEACHING
全国語学教育学会

Modeling Complexity, Accuracy, and Fluency of Japanese Learners of English: A Structural Equation Modeling Approach

Rie Koizumi
Juntendo University
Yo In'nami
Shibaura Institute of Technology

With this study, we aimed to obtain a better understanding of the factor structure of the complexity, accuracy, and fluency (CAF) of English speaking proficiency. For this purpose, 224 Japanese junior and senior high school students with an English level of elementary to lower intermediate took an English speaking test. We transcribed what they said, computed measures to assess CAF, and used structural equation modeling (SEM) to examine whether the model in which the CAF factors are related fit the data. We found that syntactic complexity (SC), accuracy, speed fluency, and repair fluency represent distinct factors and that there are weak, moderate, or strong correlations among these factors. This generally suggests that those who speak fluently by using more words per minute tend to repair their speech more, but they also produce more accurate utterances with more clauses. We suggest pedagogical implications of considering CAF separately in teaching and assessment and benefits of using SEM for analyzing CAF.

本研究では、スピーキング熟達度における複雑さ、正確さ、流暢さ（complexity, accuracy, and fluency: CAF）の因子構造を調べる。中学生・高校生（初級から中級下レベル）の日本人学習者224名に、スピーキングテストを受けてもらった。発話をCAFの指標で数値化し、CAF因子が関連しあうモデルを共分散構造分析を用いて分析した。その結

果、統語的複雑さ、正確さ、スピードに関する流暢さ、修正に関する流暢さの4因子の相関モデルがデータに適合し、4因子は関連しあいながらも別個に捉えられることが分かった。因子間の関連は弱いものから強いものがあったが、全体的には、1分間あたりにより多くの語を使って話す学習者は、修正をより多く行うが、より正確な発話と、より多くの節を産出する傾向が見られた。指導や評価の際にCAFを別々に考慮することの重要性や、共分散構造分析でCAFを分析する利点が示された。

Second language speaking proficiency and performance has garnered increasing attention from researchers into L2 learning and assessment. One way to conceptualize L2 proficiency and performance is to use the components of complexity, accuracy, and fluency (Housen & Kuiken, 2009; Housen, Kuiken, & Vedder, 2012). These three factors (or constructs), hereinafter abbreviated as CAF, have been extensively measured in numerous studies (e.g., Foster & Tavakoli, 2009; Robinson, 2001). Despite the wide use of CAF, some issues remain unresolved (Housen & Kuiken, 2009), such as how CAF can be measured and the extent to which CAF are interrelated. To deal with these two issues, we attempt to model CAF using data from L2 Japanese learners of English with English proficiencies of elementary to lower intermediate level by employing structural equation modeling (SEM). Explicit modeling employing SEM helps in understanding the nature of CAF and their measures.

Background

Although CAF are now often grouped together, it was only in the 1990s that pedagogical and research considerations of fluency and accuracy began to be combined with the concept of complexity (Housen & Kuiken, 2009). CAF are often measured using discourse analytic measures derived from quantifying target aspects in utterances and computing values that reflect a certain dimension of language use (see Ellis & Barkhuizen, 2005; Housen et al., 2012).

Complexity is commonly defined as "the ability to use a wide and varied range of sophisticated structures and vocabulary in the L2" (Housen et al., 2012, p. 2). According to Bulté and Housen (2012), complexity is subdivided into three types: propositional, discourse-interactional, and linguistic. The former two refer to the number of idea units produced and "the number and type of turn changes that learners initiate and the interactional moves and participation roles that they engage in" (Bulté & Housen, 2012, pp. 24-25). Linguistic complexity encompasses a wide range of linguistic features; it is further classified into four dimensions: lexical (words and collocations),

morphological (inflectional/derivational levels), syntactic (sentential, clausal, and phrasal levels), and phonological (segmental/suprasegmental levels). Among these, the most discussed and researched dimension is syntactic complexity (SC). The most examined (sub)factors underlying SC are overall and sentential subordination, and these are typically indexed by mean length of unit and clauses per unit (Bulté & Housen, 2012).

Accuracy refers to "the ability to produce target-like and error-free language" (Housen et al., 2012, p. 2) and is measured using global measures (e.g., the percentage of error-free clauses) or specific measures (e.g., the percentage of correct pronouns).

Fluency is defined as "the ability to produce the L2 with native-like rapidity, pausing, hesitation, or reformulation" (Housen et al., 2012, p. 2). Tavakoli and Skehan (2005) subcategorized fluency into speed fluency (measured by, for example, speech rate and mean length of run), repair fluency (assessed with measures of reformulation, repetition, false starts, and replacements), and breakdown fluency (operationalized as pause-related indices). Bosker, Pinget, Quené, Sanders, and de Jong (2013) showed that speed fluency and breakdown fluency contribute more to raters' fluency ratings than does repair fluency.

CAF Factors

Although CAF factors are assumed to be reflected in the measures used in previous CAF studies, empirical validation studies of CAF measures—that is, investigations into whether the measures indeed assess CAF factors—have been rare (Norris & Ortega, 2009; Sheppard, 2004; Wolfe-Quintero, Inagaki, & Kim, 1998). In order to explore the relationships among CAF factors and measures, L2 researchers have used exploratory factor analysis (e.g., Mehnert, 1998; Ortega, 1995) as well as simple correlations (e.g., Kormos & Dénes, 2004; Mora & Valls-Ferrer, 2012). In this study, we focus on previous studies using exploratory factor analysis because this method can explicitly handle both latent factors (i.e., underlying or unobserved) and observed variables. This is more appropriate for examining factor structures because with correlations only observed variables can be examined.

Table 1 summarizes nine factor-analytic studies that aimed to either identify redundancy among measures and select representative measures for further analysis (Nitta, 2007; Ortega, 1995) or explore relationships among measures and identify underlying structures (the remainder of the studies in Table 1). Unfortunately, except for the study done by Sakuragi (2011),

all studies had rather small sample sizes (ranging from 17 to 80), which is known to cause instability in factor structures. Consequently, it would be safer to reinterpret previous findings than to take them at face value. On the basis of their simulation studies, Guadagnoli and Velicer (1988) argued that a factor loading pattern of .60 is likely to be stable when the sample size is 150 or greater and that a factor loading of .80 tends to be stable even when the sample size is 50. Accordingly, only measures with loadings of .80 or above were used for our reinterpretation, although the method used may produce rather conservative interpretations.

Table 1. Previous Studies Analyzing the CAF Factor Structure of Speaking Proficiency

Study [*N*]	**L2; Proficiency level**	**No. of measures included**	**Extraction; Rotation method**	**Reinterpreted factors (measures with loadings of .80 or above)**
Ortega (1995) [32]	Spanish; Upper intermediate	10 (2LC, 3A, 5F)	PCA; Oblique	1. Speed fluency (words per utterance[a]; propositions per utterance[a]; unpruned syllables per second; pruned syllables per second) 2. Accuracy (percentage of correct noun-modifier agreement; percentage of correct articles)
Skehan & Foster (1997) [72]	English; Pre- intermediate	9 (1SC, 1A, 1F for 3 tasks)	PCA; Varimax	1. SC (clauses per c-unit for the first and the third tasks)
Mehnert (1998) [31]	German; Intermediate	13 (3SC, 1LC, 4A, 5F)	Not reported	1. Speed and breakdown fluency (unpruned syllables per second; pruned syllables per second; mean length of run; total pausing time) 2. Accuracy (errors per 100 words; error-free clauses per clause; number of lexical errors) 3. SC (words per c-unit; subordinate clauses per T-unit; S-nodes per T-unit)

Study [*N*]	**L2; Proficiency level**	**No. of measures included**	**Extraction; Rotation method**	**Reinterpreted factors (measures with loadings of .80 or above)**
Taki-guchi (2003) [17]	English; Elementary	31 (3SC, 6A, 11F plus 11 non-CAF mea-sures)	PCA; Varimax	1. SC (subordinate per AS-unit; clauses per AS-unit; pruned to-kens per AS-unit; pruned tokens per turn; turns per minute) 2. Speed fluency (pruned tokens per minute; unpruned tokens per minute) 3. Accuracy (errors per minute; AS-units with errors per AS-unit; errors per AS-unit; errors per pruned token; error-free clauses per clause; error-free AS-unit per AS-unit) 4. Repair fluency (AS units with disfluency markers; disfluency markers per minute; AS units with disfluency markers per AS-unit) 5. Breakdown fluency (pauses per minute; AS-units with pauses per minute)
Shep-pard (2004) [82]	English Elementary	27 (4SC, 3LC, 4A, 16F)	PCA; Varimax	1. Speed fluency (unpruned tokens per minute; pruned tokens per minute; unpruned syllables per minute; pruned syllables per minute) 2. SC (clauses per T-unit; verbs per T-unit; phrases per T-unit; pruned tokens per T-unit) 3. Breakdown fluency (3 types of percentages of pause time, with the cut-off point of 250 milliseconds, 600 milliseconds, and 1 second) 4. Accuracy (error-free clauses per clause; error-free clauses per T-unit; error-free phrase per phrase)
Skehan & Foster (2005) [61]	English; Intermedi-ate	9 (1SC, 2A, 6F)	PCA; Varimax	1. Accuracy (error-free clauses per clause; accuracy for clauses of five words or more)

Study [*N*]	L2; Proficiency level	No. of measures included	Extraction; Rotation method	Reinterpreted factors (measures with loadings of .80 or above)
Tavakoli & Skehan (2005) [80]	English; Elementary and intermediate	12 (1SC, 1A, 10F)	Not reported	1. Speed and breakdown fluency (syllables or words per minute; total amount of silence; time spent speaking; number of pauses; mean length of pause) 2. Repair fluency (number of reformulations; number of false starts)
Nitta (2007) [27]	English; Elementary and advanced	13 (3SC, 3A, 7F)	PCA; Varimax	1. Speed and breakdown fluency (total length of pauses; mean length of run; pruned tokens or syllables per minute; number of mid-clause pauses) 2. Accuracy (error-free clauses per clause; percentage of correct verb forms) 3. Others[b] (number of chaining integration devices; number of filled pauses)
Sakuragi (2011) [113]	Japanese; Intermediate and advanced	10 (2SC, 2LC, 3A, 3F)	Principal factor analysis; Promax	1. SC (clauses per AS-unit; subordinate clauses per AS unit; pruned tokens per AS-unit) 2. Accuracy (errors per clause; error-free AS-units per AS-unit; errors per AS-unit)

Note. Only studies analyzing speaking proficiency were included. Factors with two or more measures with loadings of .80 or above (rounded off) were presented. SC = syntactic complexity; LC = lexical complexity; A = accuracy; F = fluency; PCA = principal components analysis.

[a]Although Ortega originally considered the number of words per utterance and that of propositions per utterance as SC measures, she doubted the validity of such interpretation in the discussion; Norris and Ortega (2009) further interpreted the number of words per utterance as reflecting fluency in the same manner as the mean length of run. [b]Only appeared under the online planning condition.

A reinterpretation of the results of previous studies yields several findings. First, three of the studies obtained at least three factors each for SC, accuracy, and fluency (Mehnert, 1998; Sheppard, 2004; Takiguchi, 2003); two studies obtained an accuracy factor and a fluency factor (Nitta, 2007; Ortega, 1995); and one study obtained an SC factor and an accuracy factor

(Sakuragi, 2011). The others all had one factor reflecting either SC, accuracy, or fluency (e.g., Skehan & Foster, 1997, 2005; Tavakoli & Skehan, 2005). These results indicate that accuracy appeared as a factor in most cases, followed by fluency and SC, but that all three were not always present. This suggests insufficient empirical evidence about whether we can derive distinct CAF factors. Research Question 1 was designed to examine this aspect.

Second, an SC factor appeared as a single dimension in five studies (e.g., Mehnert, 1998; Sheppard, 2004). The derivation of one SC factor consistently across studies may suggest that SC dimensions—for example, overall SC and sentential-subordination SC—can be conceptually distinguished but not empirically discriminated (Pallotti, 2009). In addition, in all five studies that derived an SC factor, one measure—the number of tokens (i.e., words) divided by the number of units (e.g., T-units)—loaded on the SC factor. Only one study, Skehan and Foster (1997), did not use this measure. The interpretation of this measure, called *mean length of unit* or *unit length*, has been controversial. Although the mean length of run, or the number of syllables/tokens per unit primarily related to pause or repair, is interpreted as fluency (e.g., Segalowitz & Freed, 2004; Tavakoli & Skehan, 2005), the number of tokens per primarily syntactic unit (e.g., T-unit) has two distinct accounts: fluency (e.g., Ishikawa, 2007; Robinson, 2001; Wolfe-Quintero et al., 1998) and SC (e.g., Bulté & Housen, 2012; Koizumi, 2005b; Norris & Ortega, 2009). Results of CAF factor-analytic studies support the latter interpretation (e.g., Mehnert, 1998; Sakuragi, 2011) because the number of tokens per syntactic unit loaded on an SC factor. In addition, it was found that no lexical complexity factor emerged as distinct. This may be attributable to the limited number and types of measures of lexical complexity employed in factor-analytic studies.

The accuracy factor, if any, consistently appeared as a single dimension (e.g., Nitta, 2007; Ortega, 1995). This supports Pallotti's (2009) observation that the accuracy factor is stable in nature. Further, fluency has been found to comprise up to three factors. It appears that more factors are extracted when more relevant fluency measures are involved (e.g., 3 fluency factors using 11 fluency measures in Takiguchi, 2003). The interpretations of factors suggest that the speed dimension is often linked to the breakdown dimension. Moreover, the speed dimension tends to be a primary component of fluency, whereas the repair dimension is found as a separate factor and a secondary dimension of fluency.

Although previous factor-analytic studies have provided an invaluable foundation for clarifying CAF factors and measures, three methodological

issues must be addressed to derive stronger evidence. First, it is unclear if the data satisfied the statistical assumptions for using exploratory factor analysis. Although some studies (e.g., Ortega, 1995; Sakuragi, 2011) conducted Bartlett's test of sphericity to determine if a correlation matrix was adequate for factor analysis, no studies reported multivariate normality—another essential assumption for factor analysis: "all variables, and all linear combinations of variables, are normally distributed" (Tabachnick & Fidell, 2007, p. 613). Second—as Plonsky and Gass (2011) argued about L2 studies in general—some CAF studies (e.g., Mehnert, 1998; Tavakoli & Skehan, 2005) did not report how factor analysis was conducted, such as what extraction and rotation methods were used or how the number of factors was determined. This is troubling because results would change according to these specifications (e.g., Tabachnick & Fidell, 2007). Further, except for studies done by Ortega (1995) and Sakuragi (2011), varimax rotation was the rotation method of choice, which assumes no correlations among extracted factors. However, this is often too strong an assumption to hold because interrelationships between factors can usually be hypothesized; thus, oblique (e.g., Promax) rotation is recommended.

Third, exploratory factor analysis is of limited value due to its data-driven nature. Given a growing number of previous studies on CAF that permit the construction of a theory-based model, SEM is a more suitable method. However, thus far, no studies have employed this method for CAF analyses. The following are the main advantages of SEM (Byrne, 2006). First, SEM uses a confirmatory, hypothesis-testing method. It can model not only observed variables but also latent variables (i.e., factors) and can flexibly model complex relationships on the basis of previous findings. Second, it can separate measurement errors from observed and latent variables and estimate relationships among the variables that are being investigated, thereby statistically controlling for such errors. One source of errors is the variability caused by task differences—earlier studies suggested that task variations (e.g., cognitive demand of tasks) led to different speaking performances (e.g., Tavakoli & Skehan, 2005; Robinson, 2001). As SEM is a large-sample technique (usually requiring a sample size of at least 100), its application to the investigation of the CAF structure is rather difficult in studies in which sample size is considerably smaller. In order to take full advantage of SEM, we collected a large sample, tested statistical assumptions, and examined the CAF factor structure.

Relationships Among CAF

A second aspect investigated in this study is the question of how CAF are interrelated. Norris and Ortega (2009) indicated the need for research into revealing the interdependence and dynamism of CAF using multivariate modeling such as SEM. Generally, positive and relatively strong relationships are predicted because CAF are expected to improve gradually as learners' proficiency increases, although not necessarily simultaneously. However, previous studies have reported divergent degrees of correlations, even among learners with a wide range of proficiency.

For example, a weak correlation was reported in Sakuragi (2011), which documented the relationship between SC and accuracy (r = .19) among 113 intermediate and advanced learners of Japanese. Ortega (1995) also presented low correlations between accuracy and speed fluency (r = .08 to .22) among 32 upper intermediate learners of Spanish. Further, Koizumi (2005b) reported marginal to fairly weak correlations among SC, accuracy, speed fluency, and repair fluency (r = -.21 to .47) among 74 elementary to upper elementary Japanese learners of English. Kormos and Dénes (2004) reported a moderate correlation between accuracy and speed fluency (r_s = .66) and a low correlation between speed fluency and repair fluency (r_s = -.19) among 16 low-intermediate and advanced Hungarian learners of English. These varied correlations suggest the need for further investigation and lead to Research Question 2.

The Current Study

To clarify the CAF structure, we examine factors of SC, accuracy, and fluency (fluency is further divided into speed fluency and repair fluency), as well as the relationship among these factors. Two research questions were asked with a specific focus on Japanese learners of English at the elementary to lower intermediate level.

RQ1: Do complexity, accuracy, and fluency (CAF) represent distinct factors?

RQ2: How are complexity, accuracy, and fluency (CAF) interrelated?

Method

Participants

The participants were 224 Japanese learners of English—97 males and 127 females—attending 10 junior or senior high schools, aged from 14 to

18. Their first language was Japanese. They had received EFL instruction at secondary schools in Japan for from 2 to 5 years. The overall English proficiency levels on the Eiken Test (Society for Testing English Proficiency [STEP], 2011) were reported by the participants and ranged from Grades 5 (2%) to 2 (4%), with the majority at Grades Pre-2 or 3 (61%), although 23% reported no experience of taking the Eiken Test. According to the STEP (2011), Eiken Test Grades 2 to 5 are roughly equivalent to the A1 to B1 levels of the Common European Framework of Reference for languages (Council of Europe, 2001). Thus, the participants were considered to have novice- to lower intermediate-level English proficiency. They were selected for participation in this study from a larger sample only if they took a speaking test and produced at least one clause for every speaking task.

Instrument

The students took a speaking test that contained five tasks to elicit real-time monologues without pretask planning time (Koizumi, 2005a). The test lasted for 15 minutes in a tape-mediated format. Task 1 was a self-introduction task, Tasks 3 and 4 involved describing a single picture, and Tasks 2 and 5 involved explaining the differences between two pictures (see Appendix for a sample of utterances). We used these five tasks to tap wider areas of speaking proficiency. The output from learners was limited; the mean of the number of tokens for each task ranged from 25.39 (*SD* = 10.98) in Task 3 to 37.14 (*SD* = 13.64) in Task 1.

Analyses

We created the coding scheme using Foster, Tonkyn, and Wigglesworth's (2000) definitions. Raters (native speakers and highly proficient Japanese learners of English) practiced coding and later, using the scheme, independently coded one-third (randomly sampled) of the transcribed utterances for each task (45 seconds for each task; a total of 225 seconds) for features such as the number of AS-units. The number of raters varied depending on the coded features: Four raters were used for assessing error-free clauses because of difficulty in judgment; two raters were employed for the other features. The inter-coder reliabilities were found to be high (e.g., r = .86 to 1.00 for the number of AS-units, clauses, and disfluency markers; α = .86 to .93 for the number of error-free clauses). Because the features were measured on interval scales, we used Pearson product-moment correlations for two raters and Cronbach's alpha for four raters. Further, we resolved dis-

agreement through discussion and created the final detailed coding scheme that clarified the aspects on which raters diverged and that required little judgment from raters.

The remainder of the transcripts were coded by a single rater (the first author) who had judged one-third of the transcripts for all the coded features; she coded them while examining the coding scheme carefully. This method can be justified because the inter-coder reliability among raters for one-third of the transcripts was sufficiently high and because this is a common procedure for coding data (see Révész, 2012).

For the analysis of speaking proficiency, we computed five discourse analytic measures for each task (see Table 2). Similar measures were initially computed but excluded because of high correlations with the remaining measures (e.g., number of disfluency markers per token) and inconsistent results across tasks (number of tokens per clause). We did not include pause-based measures due to poor recording conditions that hampered such in-depth analysis.

Table 2. Summary of Five Measures

Factor	Code	Measure	Source example
Syntactic complexity (SC)	SC1	Overall SC: AS-unit length: No. of tokens per AS-unit	Mehnert (1998)
	SC2	Sentential-subordination SC: No. of clauses per AS-unit	Tavakoli & Skehan (2005)
Accuracy	A	No. of error-free clauses per clause	Skehan & Foster (2005)
Fluency	F1	Speed fluency: No. of tokens per minute	Sheppard (2004)
	F2	Repair fluency: No. of disfluency markers per minute	Sheppard (2004)

Note. Tokens (i.e., words) refers to pruned tokens after disfluency markers were excluded (i.e., functionless repetitions, self-repairs, and filled pauses, such as *mm*, *ah*). The definition of clauses was based on Foster et al. (2000; for instance, the utterance "I like reading books" had two clauses: *I like* and *reading books*). Abbreviations in this table are used in text and figures.

For SEM analyses, we used EQS (Version 6.1; Bentler, 2010), but drew diagrams using Amos (Version 7.0.0; Arbuckle, 2006) for visual display. The SEM analyses enable us to examine whether a model depicting relationships between variables that is based on a theory or the literature, or both, fits the data. If it fits the data, it implies that relationships among variables specified in the model accord well with relationships among variables in the data. Then, we can empirically interpret the findings to mean that the model represents the data well and that the data has a factor structure in which factors and observed variables are related as specified in the model.

The requisites for appropriate SEM practices (e.g., Byrne 2006; In'nami & Koizumi, 2011) involve normality, parameter estimation methods, model fit indices used, missing data treatment, and sample size. Univariate and multivariate normality of measures was judged on the basis of skewness and kurtosis values and Mardia's normalized estimate and found to be violated. Thus, the robust maximum likelihood method was used. One of the factor loadings from each factor was fixed to 1.00 for scale identification. Model fit was checked by the comparative fit index (CFI) of 0.90 or above (Arbuckle & Wothke, 1995), root mean square error of approximation (RMSEA) of 0.08 or below (Browne & Cudeck, 1993), standardized root mean square residual (SRMR) of .08 or below (Hu & Bentler, 1999), and other indices. There were no missing data. The sample size exceeded 200, which is considered large according to Kline's (2005) guidelines. Further, intervariable Pearson product-moment correlations (r = –.13 to .74) were not so high as to cause problems of multicollinearity (r = .90 or above; Tabachnick & Fidell, 2007).

Results

We followed four steps. First, we constructed a model with only one latent factor assessed by five observed variables (one measure from each task). Figure 1 depicts a model (Model 1) of speed fluency in which a factor is represented by an oval (Speed fluency), observed variables are represented by rectangles (F1 from five tasks; T1F1 [Task 1 F1] to T5F1), and measurement errors (e1 to e5) are represented by circles. One-headed arrows depict the influence of the speed fluency factor on the five variables, which are also affected by five errors. Based on existing literature, this model specifies that there are five F1 variables underlying a speed fluency factor, but that there are some aspects of F1 variables that are unexplained by the factor but explained by errors. We also constructed four other models separately (for overall SC, sentential-subordination SC, accuracy, and repair fluency) but have not included them here because of lack of space.

Second, we tested whether each model fits the data. Fit statistics for five models indicate that all models fit the data (e.g., CFI = .95 to 1.00, RMSEA = 0.00 to 0.077, SRMR = .02 to .04), thereby indicating that the measures used represented each factor well.

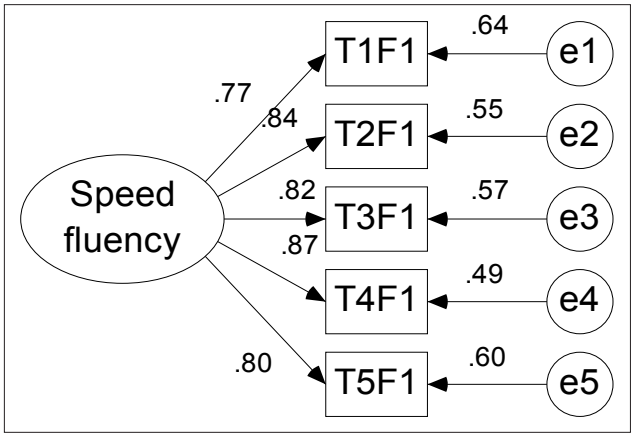

Figure 1. One-factor model for speed fluency (Model 1).
T1 = Task 1; F1 = number of tokens per minute; e = measurement error. Standardized estimates are shown. All the testable path coefficients were significant.

Third, we constructed a model (Model 2) with five CAF factors (overall SC, sentential-subordination SC, accuracy, speed fluency, and repair fluency), all of which were related to one another (this model has not been displayed in this paper due to space limitations). This model did not fit the data (e.g., CFI = .84; RMSEA = 0.07 [95% confidence interval: 0.06, 0.08]; SRMR = .07) mainly because the correlation between overall SC and sentential-subordination SC was too high (r = 1.03). We retained a factor of sentential-subordination SC because SC2 (number of clauses per AS-unit) is considered a more typical measure of SC than SC1 (number of tokens per AS-unit), which is occasionally used as a fluency measure. Because of the strong correlation between the two SC factors, the results derived from SC2 can be considered to be applicable to SC1, and the sentential-subordination SC factor is hereinafter interpreted as SC in general.

Finally, we tested a model with four factors (SC, accuracy, speed fluency, and repair fluency) that were correlated with one another (indicated by

two-headed arrows), as evident in Model 3 in Figure 2. Fit statistics of this model were sufficient (e.g., CFI = .93; RMSEA = 0.05 [0.04, 0.06]; SRMR = .06). Other competing models did not fit the data well, such as one with a higher order speaking proficiency factor represented by the four factors (e.g., CFI = .84; RMSEA = 0.08 [0.07, 0.09]; SRMR = .17) and another with a unitary speaking proficiency factor without any CAF factors (e.g., CFI = .73; RMSEA = 0.10 [0.08, 0.11]; SRMR = .10).

The standardized estimates range from –1.00 to 1.00 and are interpreted in the same manner as the correlation and regression coefficients, with values close to zero indicating marginal associations and those close to –1.00 or 1.00 indicating strong associations. A good fit of Model 3 to the data suggests two types of relationships: those between a factor and each observed variable and those among factors. First, all observed variables were shown to reflect each factor well, thereby indicating that the variables assessed each factor appropriately. Further, path coefficients from speed fluency and repair fluency factors to observed variables were found to be strong (β = .65 to .88), whereas those from accuracy and SC factors were moderate (β = .20 to .57). This indicates that there was less variation in the path coefficients for fluency than in those for SC and accuracy, thereby suggesting that fluency measures may be more generalizable across tasks.

Second, the model indicates positive but varied degrees of relationships among CAF factors: SC was more closely related to accuracy (r = .88) than to speed fluency and repair fluency (r = .63 and .43, respectively). Further, accuracy was more closely related to speed fluency (r = .35) than to repair fluency (r = .13), and the two fluency factors were moderately correlated (r = .66).

Discussion

Do CAF Represent Distinct Factors?

A good fit of Model 3 in Figure 2 suggests that the answer to the question of whether CAF represent distinct factors is affirmative. A structure with distinct CAF factors accords well with Mehnert (1998), Sheppard (2004), and Takiguchi (2003) but does not with others (e.g., Nitta, 2007; Ortega, 1995). Researchers in all previous studies used exploratory factor analysis and attempted to extract CAF factors reflected in different measures, whereas we used SEM with each factor reflected in the same measures from five tasks. The results indicate that our approach is useful for examining the distinctiveness of CAF factors.

The result that two fluency factors (speed and repair) were moderately positively associated (r = .66) is indicative of their distinct yet related na-

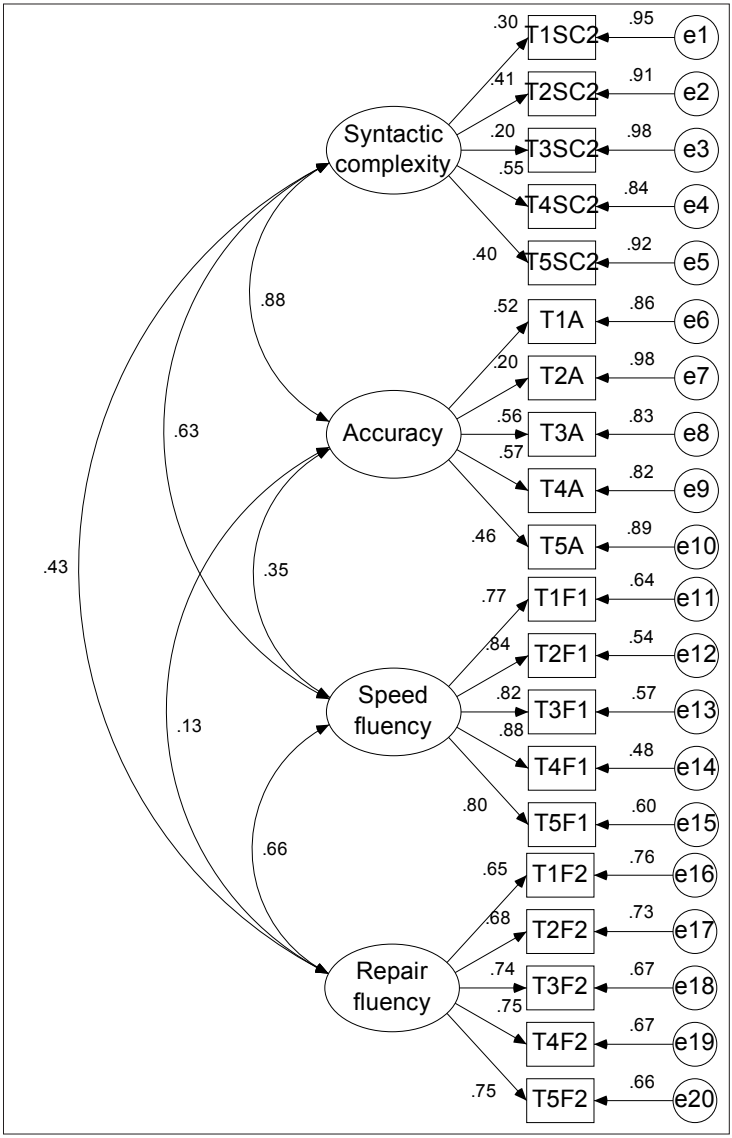

Figure 2. Four-factor correlated model for CAF (Model 3).
T1 = Task 1; F1 = number of tokens per minute; e = measurement error. Standardized estimates are shown. All the testable path coefficients were significant.

ture. This is in line with previous exploratory factor analyses that reveal the distinct characteristics of speed and repair fluency (Takiguchi, 2003; Tavakoli & Skehan, 2005), but other studies reported only negative correlations between speed and repair fluency (e.g., r = -.19 in Kormos and Dénes, 2004). The positive correlation between speed fluency and repair fluency factors in our study implies that—given that F2, which taps repair fluency, is calculated by the number of *disfluency markers* per minute—speakers who produce more tokens (excluding disfluency markers) tend to use *more* repetitions and self-repairs and produce *more* filled pauses in their utterances. This could be explained by the participants' lower proficiency levels and the targeting of a wide range of proficiency levels. It is possible that at this proficiency range, those who try to search for and utter words more rapidly are unable to avoid hesitation due to insufficient automatized skills, as reported in Wood (2010). Alternatively, learners with higher proficiency can monitor their utterances (Kormos, 2006); therefore, they repair their speech more while speaking faster. However, we also found a very weak relationship between accuracy and repair fluency (r = .13), in line with Koizumi (2005b; r = -.05 to .21), which suggests that more repairing does not likely lead to more accurate speech according to the proficiency range of the current study.

In addition, although a strong claim cannot be made due to the lack of a model fit, a very strong relationship (r = 1.03) between factors of overall SC and sentential-subordination SC in Model 2 indicates that the length of the AS-unit is an SC measure, which supports all previous studies (e.g., Norris & Ortega, 2009; Sakuragi, 2011). It also indicates that although they are differentiated conceptually, dimensions of overall SC and sentential-subordination SC are not empirically distinct among learners, or at least among learners at a lower proficiency level.

How Are CAF Interrelated?

As displayed in Figure 2, CAF were found to be independent but related to varying degrees (r = .13 to .88). Overall, the model suggests that those who try to speak fluently by using more words per minute tend to repair their speech more; however, they also produce more accurate utterances with a greater number of clauses and longer units (sentences). Further, the results also indicate that as learners progress from beginning to lower intermediate levels, they develop the ability to produce such speech, thereby gradually improving SC, accuracy, and speed fluency (although not necessarily synchronously).

There were moderate or strong positive correlations of SC with accuracy (r = .88) and speed fluency (r = .63), whereas there was a weak relationship between accuracy and speed fluency (r = .35). It is speculated that improvement in fluency may lead to enhanced SC, which may result in heightened accuracy; that is, when learners learn to speak faster, they may gradually come to use a greater number of clauses and longer units (sentences) and subsequently may produce more accurate utterances. Such correlation patterns among CAF factors were not evident in previous studies. Previous studies (e.g., Ortega, 1995; Sakuragi, 2011) generally showed similar or weaker relationships than those revealed in our results (e.g., between accuracy and speed fluency, r = .08 to .22 in Ortega, 1995 vs. r = .35 in our study). The exception is relationships between accuracy and speed fluency in Kormos and Dénes (2004; r_s = .66 vs. r = .35 in our study). The higher correlations in our study may be partially because of the different statistical methods used. These results also suggest that the strengths of relationships vary across contexts.

Conclusion

The current study showed that CAF represent distinct factors that are correlated to varying degrees among elementary to lower intermediate Japanese learners of English. This insight into the CAF factor structure using a rigorous statistical method makes several contributions to the field.

The key pedagogical implication derived from this study is that English language teachers and testers should consider CAF factors of speaking proficiency separately. In planning their curricula and speaking instructions, teachers must carefully consider which of the CAF factors they should aim to enhance and how. In analytically assessing speaking proficiency, test makers should contemplate whether and to what extent to include SC, accuracy, and fluency in their rating criteria because they are all essential elements of speaking proficiency. The manner in which practitioners use this information would vary depending on the context. Some may decide to focus on all three; others may alter aspects to emphasize across activities and tasks, classes, or assessments, thereby aiming to achieve the development and assessment of balanced speaking proficiency; others may exclude SC and focus on accuracy and fluency for the criteria, based on moderate and strong relationships of SC with accuracy and fluency. Additionally, teachers should know that at lower proficiency levels, repetitions, self-corrections, and filled pauses tend to increase along with an increase in words uttered. Given the importance of speed fluency over repair fluency (Bosker et al., 2013), teach-

ers should devote more attention to the development and assessment of speed fluency rather than repair fluency and encourage learners to speak more rather than discourage the use of words for repair.

Our results may be limited to the study context. We targeted Japanese elementary- to lower intermediate-level learners of English, using speaking tasks that elicited basic monologues and a limited number of speaking measures. Greater generalizability of results would need replication studies in different contexts, for example, by using more cognitively challenging tasks (e.g., discussions and debates). In contrast, the strengths of our study are that it includes a larger number of learners than other CAF studies and involves meticulous analyses using SEM. SEM enabled us to separate measurement errors from variables of interest in the model and conduct a more rigorous analysis of relationships in a confirmatory manner on the basis of previous studies. The following example underscores the benefits of using SEM. Accuracy and SC factors were found to be strongly correlated (r = .88), whereas simple (zero-order) correlations between accuracy (A) and SC (SC2) from the same task—when measurement error was not controlled for—were much lower (r = .02 to .37). This clearly illustrates the importance of controlling for measurement error by using SEM. Although SEM requires the use of large sample sizes, a confirmatory approach to analyzing the factor structure of CAF has helped deepen our understanding of these factors.

Acknowledgment

This research was partially based on the first author's PhD dissertation (Koizumi, 2005a) and partially supported by the Grant-in-Aid for Scientific Research (KAKENHI) of the Ministry of Education, Culture, Sports, Science and Technology in Japan (No. 22720216). We are rather grateful to Akihiko Mochizuki for his useful suggestions.

Rie Koizumi is an Associate Professor at Juntendo University. **Yo In'nami** is an Associate Professor at Shibaura Institute of Technology. Their area of interest is modeling language ability structures. The following website contains supplementary material for this article: http://www7b.biglobe.ne.jp/~koizumi/Koizumi_research.html

References

Arbuckle, J. L. (2006). Amos (Version 7.0.0) [Computer software]. Spring House, PA: Amos Development Corporation.

Arbuckle, J. L., & Wothke, W. (1995). *Amos 4.0 user's guide.* Chicago: SmallWaters Corporation.

Bentler, P. M. (2010). EQS (Version 6.1) for Windows [Computer software]. Encino, CA: Multivariate Software.

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing, 30,* 159-175. http://dx.doi.org/10.1177/0265532212455394

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21-46). Amsterdam: John Benjamins.

Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Hillsdale, NJ: Erlbaum.

Council of Europe. (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Ellis, R., & Barkhuizen, G. (2005). *Analyzing learner language.* Oxford: Oxford University Press.

Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning, 59,* 866-896. http://dx.doi.org/10.1111/j.1467-9922.2009.00528.x

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

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103,* 265-275.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30,* 461-473. http://dx.doi.org/10.1093/applin/amp048

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 1-20). Amsterdam: John Benjamins.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55. http://dx.doi.org/10.1080/10705519909540118

In'nami, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly, 8,* 250-276. http://dx.doi.org/10.1080/15434303.2011.565844

Ishikawa, T. (2007). The effect of manipulating task complexity along the [+/- Here-and-Now] dimension on L2 written narrative discourse. In M. P. Mayo García (Ed.), *Investigating tasks in formal language learning* (pp. 136-156). Clevedon, UK: Multilingual Matters.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Koizumi, R. (2005a). *Relationships between productive vocabulary knowledge and speaking performance of Japanese learners of English at the novice level* (Unpublished doctoral dissertation). University of Tsukuba, Japan. Retrieved from http://www.tulips.tsukuba.ac.jp/limedio/dlam/B25/B2599596/1.pdf

Koizumi, R. (2005b). Speaking performance measures of fluency, accuracy, syntactic complexity, and lexical complexity. *JABAET (Japan-Britain Association for English Teaching) Journal, 9,* 5-33.

Kormos, J. (2006). *Speech production and second language acquisition.* Mahwah, NJ: Erlbaum.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32,* 145-164. http://dx.doi.org/http://dx.doi.org/10.1016/j.system.2004.01.001

Mehnert, U. (1998). The effects of different length of time for planning on second language performance. *Studies in Second Language Acquisition, 20,* 83-108.

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal Instruction and study abroad learning contexts. *TESOL Quarterly, 46,* 610-641. http://dx.doi.org/10.1002/tesq.34

Nitta, R. (2007). *The focus-on-form effects of strategic and on-line planning: An analysis of Japanese oral performance and verbal reports* (Unpublished doctoral dissertation). University of Warwick, UK.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30,* 555-578. http://dx.doi.org/10.1093/applin/amp044

Ortega, L. A.-O. (1995). *Planning and second language oral performance* (Unpublished MA thesis). University of Hawai'i.

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

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning, 61,* 325-366. http://dx.doi.org/10.1111/j.1467-9922.2011.00640.x

Révész, A. (2012). Coding second language data validly and reliably. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 203-221). West Sussex, UK: Wiley-Blackwell.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22,* 27-57. http://dx.doi.org/10.1093/applin/22.1.27

Sakuragi, T. (2011). Fukuzatsusa, seikakusa, ryuchosashihyo no koseigainendatosei no kensho [The construct validity of the measures of complexity, accuracy, and fluency: Analyzing the speaking performance of learners of Japanese]. *JALT Journal, 33,* 157-174.

Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition. *Studies in Second Language Acquisition, 26,* 173-199. http://dx.doi.org/10.1017/S0272263104262027

Sheppard, C. (2004). The measurement of second language production: The validity of fluency, accuracy, and complexity. *ICU (International Christian University) Language Research Bulletin, 19,* 139-156.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1,* 185-211. http://dx.doi.org/10.1177/136216889700100302

Skehan, P., & Foster, P. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 193-216). Amsterdam: John Benjamins.

Society for Testing English Proficiency (STEP). (2011). *Eiken: Test in Practical English Proficiency.* Retrieved from http://stepeiken.org/

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Needham Heights, MA: Allyn & Bacon.

Takiguchi, H. (2003). *A study of the development of speaking skills within the framework of fluency, accuracy and complexity among Japanese EFL junior high school students* (Unpublished MA thesis). Joetsu University of Education, Niigata, Japan.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 238-273). Amsterdam: John Benjamins.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy & complexity.* Honolulu: University of Hawai'i Press.

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence and classroom applications.* London: Continuum.

Appendix

Sample of Utterances From Five Tasks

A male participant: 1st-year senior high school student studying English for 3.5 years—claimed to have Eiken Grade Pre-2.

Task 1: Self-introduction

My name is * *. I have a sister. Her name is *. My parents {are in} are normal. My friends are many in my school.

Task 2: Comparison of two pictures

The windows is opened. The door's color is blue. {There are} there is a cow. There is a tree is around. There are four windows at the house.

Task 3: Picture description

A girl is washing a cup in the kitchen. The woman help the girl to washing. There are many books on

Task 4: Picture description

A man and a girl is riding a bike by the lake. There are many trees by the lake. The weather is very good.

Task 5: Comparison of two pictures

I think :: the apple before is one. But after, the apple is half. And {the} the book is mine before. But after, the book is name jiro. Another, there

Note. * = The student said a name. { } = repetitions, self-corrections, and other functionless words uttered (words in { } were ignored in accuracy rating and token counting). :: = subordinate clause