Zhou, Z. & Li, Z. (2017). Exploring the relationship between fluency measures and speaking performance of prospective international teaching assistants. In M. O'Brien & J. Levis (Eds). *Proceedings of the 8th Pronunciation in Second Language Learning and Teaching Conference*, ISSN 2380-9566, Calgary, AB, August 2016 (pp. 176-185). Ames, IA: Iowa State University.

EXPLORING THE RELATIONSHIP BETWEEN FLUENCY MEASURES AND SPEAKING PERFORMANCE OF PROSPECTIVE INTERNATIONAL TEACHING ASSISTANTS

Ziwei Zhou, Iowa State University, USA
Zhi Li, Paragon Testing Enterprise, Canada

Previous studies suggest that L2 fluency measures can influence the evaluation of speaking performances, but with different degrees of contribution. Such relationships are still under-researched for international teaching assistants (ITA), who play important roles in undergraduate education in the higher education institutions in North America. This study focuses on 114 prospective ITAs at a large Midwestern university in the US and aims to investigate the relationships between fluency measures and their speaking performances in an in-house speaking test for ITAs. Four categories of fluency measures, namely, speed, juncture pauses as breakdown, non-juncture pauses as breakdown, and fillers, were calculated in the form of 15 variables, based on the automated results from a modified version of Quené, Persoon, and de Jong's (2010) Praat script as well as manual annotation of the speech samples. The raw holistic scores of the speaking performance were analyzed using FACETS to obtain corrected or fair scores. A multiple regression was conducted and the results indicated that average syllable duration and normalized count of juncture pauses were the most significant predictors of the corrected score. These findings are informative for ITA programs to better understand the fluency characteristics of ITAs and their contribution to speaking performances.

INTRODUCTION

Although the notion of language fluency was used interchangeably with general language proficiency at times (e.g. "She speaks fluent French"), the impression of fluent speech was generally associated with a sense of ease, motion, fluidity, and smoothness in speech (Chambers 1997; Lennon, 2000). In speech production, fluency appears to be one of the "most easily noticeable" properties that differentiates L1 and L2 speakers (Kormos, 2006, p. 154). In his seminal work, Fillmore (1979) proposed four types of fluency in terms of language competencies: 1) the ability to fill the time with talks; 2) the ability to talk coherently with reasonable and semantically "dense" sentences; 3) the ability to speak appropriately according to specific contexts; and 4) the ability to use language creatively with novelty and imagination. Similarly, Lennon (2000) provided a working definition of fluency as "the rapid smooth, accurate, lucid, and efficient translation of thought or communicative intention into language under the temporal constraint of online-processing" (p. 26). Based on the interdependent nature of fluency issues, Segalowitz (2010) argued that fluency situated itself upon "the intersection region of the subdisciplines of cognitive science" and its inquiry should require multidimensional and multidisciplinary efforts (p. xiv).

In terms of studying fluency in L2 context, the "dual approach" is usually adopted where perceived fluency scores assigned by raters to non-native speech are compared with objective measures calculated for the same speech (Cucchiarini, Strik, & Boves, 2000; Xi & Mollaun,

2006). General findings in this line of research point to a strong positive correlation between human score and fluency measures. As reviewed by Yoon (2009), most findings revealed temporal measures of fluency to have significant roles to play in fluency scores.

More recently, researchers have paid particular attention to the possible overlapping in categorizing various fluency measures and endeavored to separate these measures by setting distinct boundaries (de Jong, Groenhout, Schoonen, & Hulstijn, 2015; Kahng, 2014). For example, de Jong et al. (2015) clustered the measures into three categories: 1) speed fluency that has been characterized as the rate and density of speech delivery; 2) breakdown fluency that concerns the extent to which a continuous speech signal is interrupted; 3) repair fluency that relates to the number of corrections and repetitions present in speech (Skehan, 2009). In particular, Kahng (2014) argued that fluency measures should be chosen in a way that they explicitly represented each aspect of fluency and should not be mathematically dependent or strongly interrelated with each other. The effort to categorize aspects of fluency and disambiguate their boundaries is necessary because one can reasonably assume that "measures from the same fluency aspect might be caused by the same cognitive problems in the speech production process" (Bosker, Pinger, Quené, Sanders, & de Jong, 2012, p. 171).

Recently in the context of language testing, there has been revitalized interest in studying fluency with more emphasis on statistical rigor, better integration of advanced speech technology, and broadened scope (e.g. including more variables such as test takers' L1 background). For instance, Ginther, Dimova, and Yang (2010) used Python interface to facilitate the automated extraction of fluency measures produced by Praat. Through administering the Oral English Proficiency Test (OEPT) via computer to 150 test takers, their analyses revealed that speech rate, speech time ratio, mean length of run, and the number and length of silent pauses were significant predictors for proficiency scores. However, fluency variables alone did not distinguish adjacent levels of the OEPT scale. Bhat, Hasegawa-Johnson, and Sproat (2010) investigated signal-level fluency quantifiers in a rated speech corpus of L2 English learners. The results of their logistic regression analyses indicated that articulation rate and phonation-time ratio significantly predicted fluency level. Using mixed-effects modeling, Bosker et al. (2012) found that listeners weighed the relative importance of the perceived fluency to arrive at holistic score or overall judgement of L2 Dutch speech. This study was further extended to 53 L2 learners of Dutch of L1 English and Turkish by de Jong et al. (2015) who validated L2 average syllable duration (ASD or inversion of articulation rate) as the most useful predictor of fluency, explaining 30% of the variance in L2 proficiency. In addition, partialling out L1 variance increased the explained variance to 41%.

Informed by the current perspective to fluency issues in language testing context, this study aims to investigate the relative contribution of fluency, as measured by its speed and breakdown domain, to the multi-componential construct of speaking proficiency (de Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012). As pointed out by Ginther et al. (2010), examining the subskills such as fluency underlying holistic score can augment our understanding in the interpretation and use of test scores and provide supporting evidence for the validity of inferences with regards to test performances. Practically speaking, information obtained from such inquiry can serve to improve fluency-related descriptors in scoring scales. Moreover, the potential usefulness of automated assessment deserves further investigations due to its advantage in providing objective measures with minimum human intervention as well as low associated expenses.

In addition, this study concerns the population of international teaching assistants (ITAs) since it has been observed that ITAs' speaking proficiency in general and speaking fluency in particular appear to be far less laudable, as compared to their high attainment of content knowledge (Gorsuch, 2011; Kaufman & Brownworth, 2006). Gorsuch 2011 particularly point out that ITAs' poor fluency, including "slow speech, false starts, and particularly pauses that violate phrasal boundaries", may pose tremendous obstacles for them to meet various academic requirements such as teaching undergraduate classes. In the light of the above accounts, this study pursues the following research questions:

1) To what extent can speed and breakdown fluency measures distinguish score levels in the ITA speaking test?
2) To what extent can speed and breakdown fluency measures predict corrected test scores in the ITA speaking test?

METHODS

The speaking test and the data. The test of interest is an English speaking test for prospective international teaching assistants (ITAs) used at a large Midwestern university in the US. The speaking test consists of two main components, namely an oral proficiency interview section and a simulated mini-lecture section which comprises a 5-minute lecturing and a 3-minute question-answering. This study only focused on the second component of the test (TEACH section) because ITAs' performance of simulated lecturing is of primary concern for individual departments and the teaching performance in the form of monologue is relatively easier to extract fluency measures, compared with the interview performances.

The speaking performance is holistically evaluated by three trained raters for overall comprehensibility, the effectiveness of oral language, and listening ability, using a 300-point score band. The scores are then converted to a 4-point scale with 4 being the lowest level (not certified) and 1 the highest level (fully certified). According to the online score guide of the speaking test, fluency is explicitly or implicitly mentioned in the descriptors of the four levels. For example, a performance at level 1 (lowest level) may be characterized as using "short utterances that are filled with hesitations, pauses, self-corrections, and ineffective reformulations" whereas a level 1 performance would "show very good fluency."

Due to the small number of level 4 test-takers, we decided not to include Level 4 samples and collected 114 TEACH speech samples with 38 sampled from each of the three levels (levels 1-3) from a pool of 227 samples rated by 11 raters. Among the 114 test-takers, 78 were males and 36 were females. The major first languages included Chinese (45), Korean (9), Vietnamese (4), Hindi (4), Nepali (4), Bengali (4), and Arabic (3). They represented graduate students from three main colleges: Engineering (58), Humanities (28), and Business (11).

Considering the fact that fluency analysis is labor-intensive and time-consuming, we limited our analysis to the first 2-minute segment of each teaching monologue sample, excluding the period of silence at the beginning of the recording as well as long non-speech pauses for blackboard writing, if present. These segments were pre-processed with Audacity 2.0.5 to normalize amplitude level and to remove background noise.

Fluency measures. Both automated measures and manual annotation-based measures of fluency were employed in this study. A Praat script written by Quené, Persoon, and de Jong (2010) was then used on Praat 6.0.17 to analyze the 2-minute segments for three automated measures of the speed aspect of fluency, i.e., speech rate (number of syllable/total time), articulation rate (number of syllables/phonation time), as well as average syllable duration (ASD, phonation time/number of syllables).

The resultant textgrid files from Praat output were used as a basis for manual annotation of breakdowns with AS-unit and clausal unit (Foster, Tonkyn, & Wigglesworth, 2010) as units of analysis. Three types of breakdown were annotated: juncture pause, non-juncture pause, and fillers. The juncture pauses are the noticeable silences (whose duration is greater than 200 ms) occurring at the boundaries of clausal units as in "we are going to learn this [pause]" or sub-clausal units that can be elaborated into complete semantic unit as in "Ok. [pause] Now, let's move to …". The non-juncture pauses are the noticeable silences occurring within clausal units as in "supply chain is [pause] managed by …". The fillers include non-nasalization fillers like "uh" and "eh" and nasalization fillers like "um" and "un" either occurring separately or attached as an elongated vowel coda.

The manual annotation was carried out by the two researchers. After initial familiarization of the annotation scheme and calibration of annotation on a set of five speech samples, each researcher annotated another set of five speech segments. 10% of the samples were later double annotated to examine inter-coder reliability. The agreement for pauses was 87% and that for fillers was 72%. The disagreement was solved through discussion which prompted a round of self-check of the annotations to improve accuracy.

The combination of automated measures and manual annotation contributed a total of 15 fluency measures to reflect four major aspects of fluency, namely, speed, juncture pauses as breakdown, non-juncture pauses as breakdown, and fillers (see Table 1 in the Results section). As mentioned earlier, the speed aspect was represented with speech rate, articulation rate, and ASD. The breakdowns were characterized with count, duration, as well as mean length of two types of pauses. In addition to these variables, we conceptualized density of speech as another aspect of fluency in terms of mean length of run (total length of sounding segment/number of sounding segment), count-based ratio (ratio of count of sounding segment by count of pauses), and duration-based ratio (the ratio of duration of sounding segment by duration of pauses).

To make the variable values comparable across speech samples, normalizations are applied to count and duration variables separately. This is different from previous practice (e.g. Kahng, 2014) where both variables are normalized against total speaking time. Thus, For example, the normalized count of juncture pause is the result of number of juncture pause count divided by the number of sounding segments given by Praat script output. Likewise, the duration of juncture pause is normalized against the corresponding phonation time.

Speaking proficiency. In this study the speaking proficiency is treated as the dependent variable and multiple fluency measures are the independent variables or predictors of speaking proficiency. Like other performance assessment, raters played an important role in the speaking test through their operationalization of the rating scale and subjective evaluation of test-takers' performance, which may introduce variability of rater severity to the final ratings. To account for this potential impact, we used multifaceted Rasch model (MFRM) to estimate rater severity with

the complete date set (N=227), which was then utilized to adjust the raw ordinal scores and produce "fair scores" of interval nature on the same reporting scale (Linacre, 2009). The multifaceted Rasch model, as an extension of the Rasch model, is capable of calibrating situational factors such as rater, testing occasions, task formats, along with the traditional parameters like test-taker ability, item difficulty using a common interval scale in the unit of logit.

Data analysis. In compliance with the variable distribution requirement from Ordinary Least Square method (OLS), we started with an examination of the normality assumption for each of the independent variable. Logarithmic or square root transformation was applied to the non-normal variables (see Table 1). In addition, a series of analysis of variance (ANOVA) were conducted to compare the independent variables across proficiency levels, followed with pairwise comparisons using Tukey HSD method.

The relationship between the fair scores and the independent variables was modeled using stepwise multiple regression. The assumption of multicollinearity was examined using the correlation matrix of the independent variables and the variance inflation factor (VIF) values of each independent variable in regression models. The assumptions of residual normality, linearity, and residual variance were checked through visually examining the corresponding scatterplots. The final model was determined based on the comparison of the Akaike's Information Criterion (AIC) values and the adjusted R^2 values of the nested models as well as the theoretical soundness of the models.

RESULTS

This section reports the results of FACETS analysis, pairwise comparisons as a part of ANOVA procedures, as well as stepwise multiple regression to answer the research questions.

FACETS analysis. The mean fit statistics for the rater facet and test-taker facet were examined. The mean infit and outfit mean square values of test-taker facet were 0.79 and 0.77, respectively, whereas the mean values for the rater facet were 0.88 and 0.69. Overall, the mean fit statistics were in the range of 0.5 and 1.5 for acceptable model fit, while 1 rater and 63 test-takers (out of 227) were identified as misfit. The chi-square test value of the rater facet was statistically significant ($\chi^2 = 30.9$, $df = 9$, $p < .01$) and the rater facet did exhibit variation in rater severity with a range -2.18 to 1.17 logits and a mean of zero, which warranted the use of adjusted scores or fair scores.

Descriptive statistics and ANOVA. Table 1 contains the means and standard deviations of the fair scores and the independent variables. To save space, an additional column was inserted to the right of Table 1 to report the statistically significant pairs based on the results of pairwise comparisons from the ANOVA procedures.

As expected, the significant differences in fair scores existed across the three proficiency levels. By contrast, five out of 15 independent variables did not show significant differences across the proficiency levels: duration of juncture pauses, mean length of juncture pauses, mean length of non-juncture pauses, mean length of fillers, and mean length of run. Three independent variables showed significant differences between levels 1 and 3 only (number of non-juncture pauses, duration of non-juncture pauses, and count of fillers). In addition to the identified difference

between levels 1 and 3, three were also different between levels 2 and 3 (duration of fillers, count-based ratio, duration-based ratio), and four variables were also different between levels 1 and 2 (speech rate, articulation rate, ASD, and number of juncture pauses).

The pairwise comparison results indicate that some independent variables may be more effective in predicting speaking proficiency in this study while the utility of others may be questionable such as the mean length-based variables.

Table 1

Descriptive statistics by proficiency levels and results of pairwise comparisons

	Level 1	**Level 2**	**Level 3**	**Pairwise comparison [a]**
fair score	1.11 (0.08)	1.99 (0.35)	2.81 (0.35)	1-2, 1-3, 2-3
speech rate	3.09 (0.46)	2.84 (0.57)	2.68 (0.38)	1-2, 1-3
articulation rate (log) [b]	1.39 (0.15)	1.30 (0.17)	1.26 (0.10)	1-2, 1-3
ASD (log)	-1.39 (0.15)	-1.30 (0.17)	-1.26 (0.10)	1-2, 1-3
count of juncture pause	0.48 (0.10)	0.42 (0.10)	0.38 (0.08)	1-2, 1-3
duration of juncture pause (log)	-1.82 (0.33)	-1.86 (0.45)	-1.80 (0.36)	none
mean length of juncture pause (log)	-0.53 (0.23)	-0.48 (0.26)	-0.39 (0.28)	none
count of non-juncture pause	0.49 (0.10)	0.54 (0.14)	0.58 (0.13)	1-3
duration of non-juncture pause (log)	-2.04 (0.38)	-1.90 (0.71)	-1.68 (0.50)	1-3
mean length of non-juncture pause	0.47 (0.09)	0.48 (0.10)	0.52 (0.10)	none
count of fillers (sqrt)	0.38 (0.17)	0.43 (0.17)	0.52 (0.17)	1-3
duration of fillers (sqrt)	0.19 (0.09)	0.21 (0.09)	0.27 (0.10)	1-3, 2-3
mean length of fillers (log)	-0.90 (0.22)	-0.99 (0.17)	-0.90 (0.17)	none
mean length of run (log)	0.52 (0.19)	0.48 (0.30)	0.43 (0.19)	none
count-based ratio	0.89 (0.10)	0.87 (0.10)	0.80 (0.10)	1-3, 2-3
duration-based ratio (log)	1.08 (0.29)	0.99 (0.46)	0.81 (0.32)	1-3, 2-3

Note: a. The pairwise comparison column shows statistically significant pairs only. b. log = logarithmic transformation, sqrt = square root transformation

Multiple regression. The correlation matrix of the independent variables indicates two pairs of highly correlated variables (over .9): articulation rate and ASD, duration of non-juncture pause and duration-based ratio. To avoid including the variables measuring similar constructs, we removed articulation rate and duration-based ratio, which left 13 variables for multiple regression analysis.

A final model was determined based on the comparison of AIC values and adjusted R^2 values of the nested models. Among the remaining 13 variables, only five variables were retained in the final model with two being significant predictors, i.e., ASD and number of juncture pause (see Table 2). There were no multicollinearity issues associated with the variables. The multiple R^2 value of the final model was .20, which means that about 20% of the fair score variance can be explained with the final model.

Table 2.

Predictor variables in the final model

Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	5.024	0.732	6.866	<.001 ***
ASD (log)	0.947	0.476	1.988	.049 *
count of juncture pause	-1.750	0.866	-2.022	.046 *
mean length of juncture pause (log)	0.392	0.313	1.252	.213
duration of non-juncture pause	-0.112	0.172	-0.649	.518
count-based ratio	-1.269	0.692	-1.834	.070

Residual SE: 0.6948 (103), Multiple R^2: 0.2012, Adjusted R^2: 0.1625, F-statistic: 5.19 (5, 103), $p < .001$

The regression coefficient of the log-transformed ASD (average syllable duration) was 0.947. In other words, a 10% increase in the log-transformed ASD will bring about an increase of 0.039 of fair scores (0.9471*log(1.1) = 0.039) or to lower the speaking proficiency level (level 1 is the highest and level 4 is the lowest level in the test) with other variables being held constant. The regression coefficient of the normalized number of juncture pauses was -1.75. This suggests that with an increase of 0.1 unit of normalized number of juncture pauses (number of juncture pause/number of sounding segments), the fair score will decrease by 0.175 points with other variables remaining unchanged. In the final model, the variable count-based ratio had a close-to-significant-level p-value (0.070) and its regression coefficient was -1.26. Similar to the variable normalized number of juncture pauses, this density variable had a negative impact on the fair score.

CONCLUSIONS AND DISCUSSIONS

Our findings about the relationships between fluency measures and speaking proficiency are, to some extent, in line with those from previous studies. In this study the importance of average syllable duration was shown in its role as a significant predictor as well as its capability of distinguishing proficiency levels 1 and 3. Similar contribution of the speed measure was also reported in de Jong et al. (2015). Other studies have highlighted pauses as important indicator of fluency, but very few made distinction between different types of pauses as this study did. Our findings about the count of juncture pauses as significant predictor confirmed that proper pauses like pauses at the boundary of grammatical units can be positively perceived. It should be noted that the explanatory power of the final model is limited partly because fluency is only one aspect of the speaking construct as embodied by the proficiency levels. Nevertheless, the findings can be used to inform both rater training for the speaking test as well as ITA instructions so that fluency can be better operationalized and taught.

Methodologically, this study covered a variety of fluency measures yielded from a combined approach with automated analyses and manual coding. Of course, there are other measures which could help predict speaking proficiency levels, for example, repair as a breakdown measure and pauses at phrasal boundaries. Future studies should take them into account as well. In addition, different modeling methods should be tried, including automated speech recognition (ASR)-based acoustic modeling and corpus-based language modeling.

ABOUT THE AUTHORS

Ziwei Zhou is a PhD student in Applied Linguistics and Technology from Iowa State University. He holds a Master's degree in TESOL from the University of Pennsylvania. His research interests include automated assessment of speaking proficiency, pronunciation learning and testing, speech processing, and natural language processing. He has presented his work in Pronunciation for Second Language and Learning conference and Midwestern Association of Language Tester conference.

Zhi Li is a Language Assessment Specialist at Paragon Testing Enterprises, BC, Canada. He holds a PhD degree in the applied linguistics and technology from Iowa State University. His research interests include language assessment, computer-assisted language learning, corpus linguistics, and systemic functional linguistics. He has presented his work at a number of professional conferences such as AAAL, LTRC, and TESOL. His research papers have been published in *System* and *Language Learning &Technology*.

REFERENCES

Bhat, S., Hasegawa-Johnson, M., & Sproat, R. (2010). Automatic fluency assessment by signal-level measurement of spontaneous speech. In *2010 INTERSPEECH Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology*, Makuhari, Japan.

Bosker, H. R., Pinget, A.-F., Quene, H., Sanders, T., & de Jong, N. H. (2012). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159–175.

Chambers, F. (1997). What do we mean by fluency? *System*, *25*(4), 535–544.

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America*, *107*(2), 989–99.

de Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(1), 5–34.

de Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*(2), 223–243.

Fillmore, L. W. (1979). Individual differences in second language acquisition. In C. Fillmore, D. Kempler, & W. Y. S. Wang (Eds.), *Individual differences in language ability and language behavior.* New York, NY: Academic Press.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375.

Ginther, A., Dimova, S., & Yang, R. (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.

Gorsuch, G. J. (2011). Improving speaking fluency for international teaching assistants by increasing input. *TESL-EJ*, *6*(4), 1–25.

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, *64*(4), 809–854.

Kaufman, D., & Brownworth, B. (2006). *Professional development of international teaching assistants*. Alexandria, VA: TESOL.

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

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach. (Ed.), *Perspectives on fluency* (pp. 25-42). Ann Arbor, MI: University of Michigan Press.

Linacre, J. M. (2009). *FACETS Rasch-model computer program*. Chicago, IL: Winsteps.com.

Quené, H., Persoon, I., & de Jong, N. (2010). *Praat Script Syllable Nuclei v2*. Retrieved from https://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei-v2

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York, NY: Routledge.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532.

Xi, X., & Mollaun, P. (2006). Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST). *ETS Research Report Series,* 2006(1), i-71.

Yoon, S.-Y. (2009). *Automated assessment of speech fluency for L2 English learners.* Unpublished doctoral dissertation. University of Illinois at Urbana Champaign.