

Revisiting the assessment potential of read-aloud speech performance: Cognitive validity and predictive validity

Shungo Suzuki, Ryuki Matsuura, Mao Saeki, and Yoichi Matsuyama
Waseda University

A read-aloud (RA) speech task has been employed in second language (L2) assessment contexts. For instance, in the Pearson Test of English (PTE) Academic, test-taker's RA speech is evaluated in terms of content (insertion/deletion of words), fluency, and pronunciation. One advantage of the use of RA speech in the assessment of L2 speaking is that due to its controlled nature by the predefined script, automated scoring can be robustly applied especially in the domain of fluency and pronunciation with the assistance of automated speech recognition and automatic extraction of temporal features (e.g., De Jong et al., 2021; Detey et al., 2020). In addition, automated scoring helps L2 learners to obtain their assessment without professional raters. If the score report of such automated scoring is meaningful to them, some positive washback effects on speech learning could be expected. However, there are two major issues in the use of RA speech performance in the assessment of L2 speaking. First, RA speech should be discussed in terms of the extent to which the cognitive processes elicited in the RA task resemble those in spontaneous speech tasks (i.e., cognitive validity; Field, 2011). Second, it is essential to understand how test-taker's RA speech can reflect their speaking performance in spontaneous speaking tasks which is close to the authentic situation of L2 speaking (i.e., predictive validity). In order to discuss the cognitive and predictive validity of RA in speaking assessment, the current study addresses the following two research questions: (a) to what extent linguistic knowledge needed in spontaneous speech production is reflected in RA speech; and (b) to what extent RA speech performance can predict spontaneous speech performance.

A total of 128 Japanese learners of English were recruited at a private university in Japan. They performed four spontaneous speech tasks and one RA task from the Speech Accent Archive (Weinberger, 2011). Spontaneous speech was evaluated by two raters in terms of fluency and phonology, using a 9-point scale. Meanwhile, following the PTE academic RA speech was analyzed by a set of automatically calculable indices capturing content, fluency and pronunciation. Cognitive processes in spontaneous speech production were operationalized based on the mechanism of speech production (Kormos, 2006; Levelt, 1989; Segalowitz, 2010). Accordingly, students' lexical, morphosyntactic, and phonological knowledge was assessed using a range of psycholinguistic tests (e.g., picture naming task).

The results of regression analyses showed that the indices of RA speech were associated with all of the lexical, morphosyntactic, and phonological measures with medium-to-strong effect sizes (Adjusted $R^2 = 13.5\text{--}48.4\%$). Meanwhile, the ratings of spontaneous speech suggested mixed validity evidence. The fluency ratings of spontaneous speech were predicted from the RA speech indices in mixed-effects modelling (Marginal $R^2 = 17.8\%$), while the phonology ratings were not predicted from either of the RA speech indices. The findings suggest that RA speech performance may offer optimal cognitive validity in relation to speech production mechanisms, while the predictive validity needs to be carefully discussed in terms of assessment criteria.

(250-word summary)

One advantage of the use of read-aloud (RA) speech in second language (L2) speaking assessment is that due to its controlled nature, automated scoring of fluency and pronunciation can be robustly achieved (e.g., Detey et al., 2020). However, there are two major issues that need to be discussed: (a) to what extent the cognitive processes in the RA task resemble those in spontaneous speech tasks (i.e., cognitive validity; Field, 2011); (b) to what extent RA speech performance can predict spontaneous speech performance (i.e., predictive validity). The current study aims to address these two validity issues of RA speech.

A total of 128 Japanese learners of English performed four spontaneous speech tasks and one RA task. Spontaneous speech was evaluated in terms of fluency and phonology, using a 9-point scale. RA speech was analyzed by a set of automatically calculable indices capturing content, fluency and pronunciation. Cognitive processes underlying spontaneous speech production were measured through a range of psycholinguistic tests tapping into students' lexical, morphosyntactic, and phonological knowledge.

The results showed that the indices of RA speech were associated with all of the lexical, morphosyntactic, and phonological measures ($R^2 = 13.5\text{--}48.4\%$). Meanwhile, only the fluency ratings of spontaneous speech were predicted from the RA speech indices ($R^2 = 17.8\%$). The findings suggest that RA speech performance may offer optimal cognitive validity in relation to speech production mechanisms, while the predictive validity needs to be carefully discussed in terms of assessment criteria.