GLODAL

OPEN ACCESS

Brief Reports in Language Research

Brief Reports in Language Research, 1, 260104 (2026)
https://doi.org/10.65553/BRLR.260104

Detecting Study Abroad Effects with an Automated Speaking Test: An Exploratory Comparison-Group Study

Ken Urano*

Faculty of Business Administration, Hokkai-Gakuen University, Sapporo, Japan

Abstract

Short-term study abroad programs are expected to demonstrate language gains, but such gains are small and hard to attribute to the program. This exploratory study used an automated speaking test, AISATS (AI Skill Assessment and Training System), and a comparison group to evaluate a three-week, speaking-focused program in the United States. Twelve Japanese university students who joined the program and nine who stayed in Japan took the test before and after it. Group × Time interactions for the total score and five component scores were estimated as standardized differences in gains (Cohen's *d*) and checked with mixed ANOVAs, analyses of covariance, and re-analyses using post-test scores produced under the test provider's updated rubric. Both groups gained similarly on the total score, largely because Task Achievement rose by more than one point in both groups, a change that is unlikely to reflect the program. With Task Achievement set aside, Fluency was the only component that favored the study abroad group (d = 0.87, 95% CI [-0.04, 1.77]; adjusted d = 0.57), although no interaction was statistically significant. Automated fluency scores may be sensitive to a short, speaking-focused stay abroad, provided that component scores are read separately and against a comparison group.

Keywords: short-term study abroad, automated speaking assessment, oral fluency, retest effect, Japanese EFL

Submission date: September 27, 2026; Revision date: October 30, 2026; Online: November 13, 2026

Copyright © 2026
The authors

(corresponding author: urano@hgu.jp)

Introduction

Universities that run overseas programs are increasingly expected to show that those programs achieve their language aims. Because a program runs every year with a new cohort, its evaluation has to be repeatable and easy to administer. A convincing account of a program's effect must also separate it from the other changes that learners undergo over the same period. Both requirements are hard to meet for short-term programs, whose gains are usually modest. In a meta-analysis of 31 studies of Japanese university learners, Hirai (2018) found that gains from stays of one month or less were much smaller than those from longer stays. Detecting a small gain and attributing it to the program requires an instrument sensitive enough to register the change and a design that can separate program effects from other developments.

Earlier studies that applied objective speaking measures to Japanese short-term programs share one limitation. Saito (2025) reported a rise in Versant speaking scores after a one-month program in Australia, and Sekiya et al. (2018) found large gains on a rated group-oral test after one-month programs in the United States, most of which had faded ten months later. Neither study included a comparison group, so the gains cannot be distinguished from changes that would have occurred without the program, concurrent coursework, or familiarity with the test. When the expected effect is small, the parameter of interest is the Group $\times$ Time interaction, that is, how much more one group changes than the other over the same interval.

Human-rated speaking tests, such as oral proficiency interviews and paired interviews, require trained raters and individual scheduling, are costly to run for every student each year, and are open to variation in rater severity and fatigue, which makes annual repeated use impractical for most programs. Automated speaking tests, which elicit and score spoken responses without a human rater, are well suited to repeated evaluation. Their composite scores correlate highly with oral proficiency interview scores (Bernstein et al., 2010), and some of them, such as Versant, define speaking as the ability to process and produce speech efficiently in real time (Van Moere, 2012). Repeated administration, however, invites retest effects. Hausknecht et al. (2007) found an average gain of about a quarter of a standard deviation when cognitive ability tests were retaken, and a test taker who has become familiar with an on-screen interviewer may respond more readily and at greater length the second time. A change in the scoring system between administrations is a further possible source of change unrelated to learning. In Messick's (1989) terms, such changes are construct-irrelevant variance. A comparison group that takes the same test over the same months absorbs them, leaving the interaction to capture the change attributable to the program.

Whether the composite or the components are read also matters. Validation studies usually report results for the composite, and the scoring algorithms are proprietary. If a construct-irrelevant change falls on a component that carries weight in the composite, the composite can move for reasons unrelated to the program and hide a smaller group difference in another component. A component-level analysis is therefore needed in addition to the comparison group.

Among the components that such tests report, fluency has the clearest link to intensive second language use. Automated fluency scores are typically based on speech rate and pausing, which correspond to the speed and breakdown dimensions of utterance fluency (Segalowitz, 2010; Tavakoli & Skehan, 2005). These are the aspects of speaking that study abroad has most consistently been found to develop. Learners who spent a semester or three months abroad came to speak faster and with fewer pauses (Mora & Valls-Ferrer, 2012; Segalowitz & Freed, 2004), whereas accuracy improved only moderately and complexity did not change in the former study, and a systematic review identified fluency gains as one of the more consistent

findings in this literature (Moore et al., 2021). Llanes and Muñoz (2009) reported measurable fluency and accuracy gains after only three to four weeks, although their 24 participants, aged 13 to 22, had no comparison group. A short program that places heavy demands on speaking might therefore leave its clearest trace on fluency rather than on knowledge-based components such as vocabulary and grammar.

The present study analyzes automated speaking test data for one cohort of a three-week program, evaluated against a comparison group, a combination that has seldom been reported for short-term programs in the Japanese context. The data were collected for the program's routine evaluation rather than for a pre-planned study, and the sample is small, so the analysis is exploratory. The expectation that Fluency would be the most program-sensitive component was formed before the analysis, but it was not preregistered, and no confirmatory test is claimed. The two research questions below ask, together, whether component scores can detect short-term gains that the composite conceals:

RQ1: Did the Study Abroad group improve more than the Comparison group on the total score of the automated speaking test?

RQ2: Did any component score, and in particular Fluency, detect a larger gain for the Study Abroad group than the total score did?

Methods

Participants and Program

The participants were 21 undergraduate business students at a private university in Japan, aged 19 to 21. Twelve (three men and nine women) took part in a three-week overseas program and formed the Study Abroad group. The other nine (six men and three women) stayed in Japan and formed the Comparison group. Students chose whether to join the program, so the groups were not randomly formed. No independent proficiency measure was available, and baseline equivalence was examined with the pre-test scores. In the first term (April to July), six of the 12 Study Abroad students and all nine Comparison students took the same elective task-based English course, which focused on oral skills. In line with task-based language teaching principles, the course prioritized meaningful oral communication and learner-driven language use over isolated structural drills (Bui, 2026a). No university English classes were held during the summer break.

The study-abroad program ran from late August to mid-September at a university in the United States. On weekday afternoons, the students attended an intensive communicative English course in small classes built around speaking, discussion, and presentations. They also visited two local companies, where they gave prepared presentations in English to business professionals and received feedback. They lived with host families throughout the stay.

Instrument

Spoken proficiency was measured with AISATS (Potential Plus, 2025), an automated speaking test that combines the Speechace scoring engine with a browser-based platform in which an on-screen AI interviewer delivers the test in about 10 to 15 minutes. The form used here contained two read-aloud items, one picture description, and two opinion questions. The test reports a total score and five component scores on a 9-point scale: Pronunciation, Fluency, Vocabulary, Grammar, and Task Achievement, the last of which concerns the relevance and completeness of responses. Fluency is computed from speaking rate and the number of pauses. The algorithms for the other components are not publicly documented. Koizumi and Nagamine

(2025) summarized the published validity information for the test: scores agreed with experienced IELTS examiners within 0.5 points for 77% of responses, and the test-retest correlation was .80. The report does not state whether these findings concern the total score or the components, or which population and tasks they came from.

Between the two administrations, the vendor changed the way scores were reported. Pre-test scores were reported on an IELTS-referenced rubric in steps of 0.5 points (total scores also included decimals). Post-test performances were scored on a new Speechace-referenced rubric in steps of 0.1 points, and the vendor also supplied post-test scores converted to the original rubric by an undisclosed procedure. The main analyses use the converted scores, and the Speechace-referenced scores serve as a sensitivity check. The two versions of the post-test scores were closely related (r = .91 to .99 across the six measures).

Procedure

Both groups took the test in May and again in October as part of the program's routine evaluation, so the interval between tests was considerably longer than the program itself. On each occasion, the students took the test individually within a designated period, in principle on a computer with a headset and without a proctor. All participants gave written informed consent for the use of their anonymized scores in research and publication. They were informed that participation was voluntary and would not affect their grades.

Analysis

The main parameter was the Group × Time interaction for each measure, expressed as Cohen's d: the difference between the groups' mean gains divided by the square root of the unweighted mean of their gain-score variances, with approximate 95% confidence intervals (normal approximation, with no small-sample correction). Positive values favor the Study Abroad group. Effect sizes were interpreted in light of their confidence intervals rather than against fixed benchmarks (Plonsky & Oswald, 2014). With this sample, significance tests have little power, so the estimates and intervals were treated as the primary result. Because the analysis is exploratory and covers six measures, no correction for multiple comparisons was applied, and the components are read as a pattern of effect sizes.

Three supplementary analyses were conducted. A 2 × 2 mixed ANOVA was run for each measure, and within-group change was tested with paired t tests. Because gain scores can reflect regression toward the mean, an ANCOVA with the pre-test score as covariate was also run, and the adjusted group difference was divided by the pooled post-test standard deviation to give an adjusted d. All analyses were repeated with the Speechace-referenced post-test scores, that is, the scores produced under the provider's updated rubric before the conversion described above. Analyses were conducted in R (R Core Team, 2025).

Results

No baseline difference was statistically significant (Welch tests, all ps ≥ .27), but the Study Abroad group began somewhat lower on Fluency (d = −0.48) and Pronunciation (d = −0.40) and higher on Grammar (d = 0.41). Table 1 reports the pre- and post-test means, and Table 2 the interaction effect sizes.

Table 1. *Pre- and Post-Test Means (and Standard Deviations) by Group*

Measure	SA Pre	SA Post	SA Gain	CG Pre	CG Post	CG Gain
Total	4.72 (0.63)	5.21 (0.40)	+0.48 (0.61)	4.90 (0.45)	5.44 (0.46)	+0.54 (0.40)
Pronunciation	5.29 (0.58)	5.54 (0.33)	+0.25 (0.50)	5.50 (0.43)	5.61 (0.22)	+0.11 (0.42)
Fluency	5.12 (0.48)	5.38 (0.31)	+0.25 (0.40)	5.33 (0.35)	5.28 (0.36)	−0.06 (0.30)
Vocabulary	5.46 (0.66)	5.25 (0.34)	−0.21 (0.66)	5.39 (0.49)	5.22 (0.51)	−0.17 (0.79)
Grammar	5.58 (0.36)	5.58 (0.36)	0.00 (0.43)	5.44 (0.30)	5.67 (0.35)	+0.22 (0.36)
Task Achievement	2.71 (0.81)	4.00 (0.85)	+1.29 (0.84)	2.83 (1.00)	4.44 (1.04)	+1.61 (0.93)

Note. SA = Study Abroad (n = 12); CG = Comparison (n = 9). Post-test scores are those converted by the test provider to the pre-test rubric (see Methods). Gains are computed from unrounded means; standard deviations of gains are in parentheses.

Both groups gained almost the same amount on the total score, and both gains were significant (Study Abroad: $t(11) = 2.74$, $p = .019$; Comparison: $t(8) = 4.11$, $p = .003$). The main effect of time was large, $F(1, 19) = 19.26$, $p < .001$, $\eta^2_G = .224$, whereas the interaction was negligible ($d = -0.12$). On the total score alone, the program would appear to have had no effect.

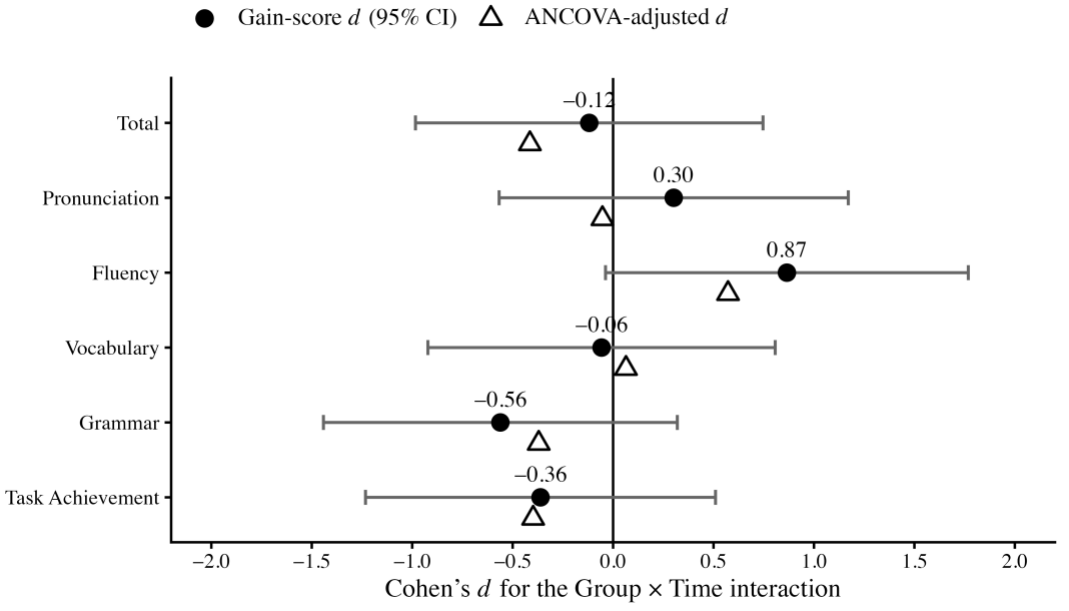
The shared gain came mostly from Task Achievement, which rose by more than one scale point in both groups, far more than any other component. Its time effect was $F(1, 19) = 56.32$, $p < .001$, $\eta^2_G = .404$, and its interaction slightly favored the Comparison group ($d = -0.36$).

Table 2. Group × Time Interaction by Measure

Measure	*d* [95% CI]	F(1, 19)	*p*	Adjusted *d*	*p*	*d* (Speechace)
Total	−0.12 [−0.98, 0.75]	0.07	.797	−0.41	.326	−0.09
Pronunciation	0.30 [−0.57, 1.17]	0.46	.508	−0.05	.898	0.34
Fluency	0.87 [−0.04, 1.77]	3.69	.070	0.57	.149	0.73
Vocabulary	−0.06 [−0.92, 0.81]	0.02	.896	0.06	.889	−0.16
Grammar	−0.56 [−1.44, 0.32]	1.58	.224	−0.37	.409	−0.72
Task Achievement	−0.36 [−1.23, 0.51]	0.68	.419	−0.40	.311	−0.38

Note. d = difference in mean gains divided by the SD of the gain scores (the square root of the unweighted mean of the groups' gain-score variances); positive values favor the Study Abroad group. F and the first p are for the interaction term of the mixed ANOVA. Adjusted d and the second p are from the ANCOVA. d (Speechace) = gain-score d with Speechace-referenced post-test scores.

Once Task Achievement is set aside, a component-specific pattern emerges (Figure 1). Fluency showed the largest interaction in favor of the Study Abroad group, $d = 0.87$, 95% CI [−0.04, 1.77]. The Study Abroad group rose by 0.25 points, $t(11) = 2.17$, $p = .053$, while the Comparison group changed little, $t(8) = -0.55$, $p = .594$. Six of the 12 Study Abroad students gained on Fluency and one declined, whereas one of the nine Comparison students gained and two declined, so the difference was not produced by one or two exceptional learners. Pronunciation moved in the same direction but less strongly. Grammar favored the Comparison group ($d = -0.56$), and Vocabulary declined slightly in both groups. No interaction reached significance, and every confidence interval included zero.



Note. Filled circles show gain-score *d* with approximate 95% confidence intervals; open triangles show ANCOVA-adjusted *d*.

Figure 1. Group × Time Interaction Effect Sizes by Measure

Adjusting for pre-test scores reduced the Fluency difference but did not reverse it (adjusted difference = 0.19 points, 95% CI [−0.07, 0.46], adjusted $d = 0.57$). The Grammar difference was likewise reduced, and the adjusted Pronunciation difference fell to nearly zero. The same pattern appeared with the Speechace-referenced scores: Task Achievement again rose by more than one point in both groups, and Fluency again showed the largest interaction in favor of the Study Abroad group, $d = 0.73$, 95% CI [−0.16, 1.62]. Finally, post-test Fluency and Grammar (converted scores) were positively correlated across participants, $r = .61$, 95% CI [.25, .83], which gives no sign that fluency was gained at the expense of accuracy. Because only half of the Study Abroad students had taken the first-term oral-skills course, the group was also split by participation (six students each). Fluency gains were +0.42 for those who had taken the course and +0.08 for those who had not, against −0.06 in the Comparison group ($d = 1.39$ and 0.41, respectively), whereas Grammar gains were 0.00 in both subgroups, against +0.22 in the Comparison group. Because every Comparison student had taken the course, participation alone cannot account for the Fluency difference, but the subgroups are too small for inference.

Discussion and conclusion

This exploratory study evaluated a three-week, speaking-focused overseas program with an automated speaking test and a comparison group. Two findings stand out. The first is a large Task Achievement gain shared by both groups. The Comparison group, which received no English instruction during the summer, gained slightly more than the Study Abroad group, whereas a program effect would predict the reverse. This change therefore tells us nothing about the program. Two explanations are available, and the present data cannot distinguish between

them. Students who had already encountered the on-screen interviewer may have responded more readily and at greater length the second time (Hausknecht et al., 2007), and a component that judges the completeness of responses may be particularly sensitive to response length. Indeed, recent exploratory evidence suggests that interacting with an AI interlocutor rather than a human partner can reduce face-threat and communicative anxiety, fostering greater willingness to elaborate in East Asian EFL contexts (Bui, 2026b). Alternatively, the change in the scoring rubric between the administrations may have shifted the Task Achievement scale, although the rise was equally large before conversion. Under either explanation the change is construct-irrelevant in Messick's (1989) sense. Its practical consequence is that one component dominated the composite. An evaluation based only on the total score would have concluded that the program had no measurable effect, and a single-group evaluation would have concluded that the participants had improved substantially. The comparison group exposed the second error, and the component-level analysis exposed the first.

The second finding is that Fluency was the only component that separated the groups in the predicted direction. The difference became smaller after adjustment for baseline scores but kept its direction, and it appeared on both post-test score scales. It is consistent with research in which oral fluency is among the most consistently reported gains from study abroad (Mora & Valls-Ferrer, 2012; Moore et al., 2021; Segalowitz & Freed, 2004), and it extends that contrast to a three-week program measured automatically (cf. Llanes & Muñoz, 2009). Theoretically, the pattern fits the view that sustained real-time use of a second language first affects the speed and smoothness of delivery, which a score based on speech rate and pausing captures, whereas knowledge-based components such as Vocabulary and Grammar are less likely to change in three weeks, although accuracy gains have been reported after stays that short (Llanes & Muñoz, 2009). The program kept speaking in constant demand through classes, prepared presentations, and homestay interaction, and the summer, when only the Study Abroad group received English instruction, was the main systematic difference between the groups. Two qualifications apply. Part of the gain-score difference probably reflects regression toward the mean, because the Study Abroad group began lower on Fluency. The Fluency score also aggregates read-aloud and free-response items, although fluency gains are expected mainly in unplanned speech (Tavakoli & Skehan, 2005), so it may dilute a program-related change or register changes in reading speed. The source of the Comparison group's small Grammar gain is unclear, since the shared first-term course focused on oral skills. The Fluency gain was concentrated among Study Abroad students who had also taken that course (+0.42 versus +0.08), so it may reflect the course and the program together.

Three practical points follow for evaluators who use automated speaking tests in pre–post designs. Component scores should be inspected routinely, because the composite can conceal both a construct-irrelevant change and a component-specific difference. Components that may be sensitive to response length are vulnerable to retest effects, so students should be familiarized with the test format before the baseline administration, and a comparison group should be included wherever possible. The scoring system of a commercial test can also change during an evaluation, so evaluators should record which version and score scale were used at each administration and request documentation of any conversion.

The limitations are substantial, and the results should be regarded as preliminary. The sample was small, the confidence intervals were wide, and no interaction was statistically significant, although a design with 12 and 9 participants has little power to detect effects of this size. Group membership was self-selected, and the groups differed in gender composition and in baseline Fluency. The testing interval extended well beyond the program and overlapped the first term, in which all of the Comparison students but only half of the Study Abroad students took the

shared oral-skills course. The test was taken without a proctor, the equivalence of the pre- and post-test scales cannot be verified, and the validity of the component scores for this population has not been established. Future work should validate the component scores against trained human raters, record the audio so that fluency can be measured separately for read-aloud and free-response items, and repeat the design with later cohorts.

In conclusion, an automated fluency score showed a difference favoring the program group that is in line with what a speaking-focused stay abroad would be expected to develop, but that awaits confirmation. Automated speaking tests can support the routine evaluation of short-term programs, provided that component scores are inspected separately, that a comparison group is used to separate program effects from changes common to all test takers, and that changes in the scoring system are documented.

Declaration of AI Use

The author used Claude (Anthropic) for language editing and for assistance in writing the R scripts. The author reviewed and verified all text, code, and output and takes full responsibility for the content.

References

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*(3), 355–377. https://doi.org/10.1177/0265532210364404

Bui, G. (2026a). Task-based language teaching: Principles and common misunderstandings. *Conceptual Foundations in Language Research*, *1*, 260101. https://doi.org/10.65553/CFLR.260101

Bui, G. (2026b). Genuine vs. virtual social presence in AI-assisted task-based interaction: An exploratory study on willingness to communicate and L2 performance across task complexity. *Brief Reports in Language Research*, *1*, 260103. https://doi.org/10.65553/BRLR.260103

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*(2), 373–385. https://doi.org/10.1037/0021-9010.92.2.373

Hirai, A. (2018). The effects of study abroad duration and predeparture proficiency on the L2 proficiency of Japanese university students: A meta-analysis approach. *JLTA Journal*, *21*, 102–123. https://doi.org/10.20622/jltajournal.21.0_102

Koizumi, R., & Nagamine, Y. (2025). Gakushū shien ni muketa AI saiten-gata supīkingu tesuto no katsuyō [Utilizing an AI-scored speaking test for learning support]. *JLTA Journal*, *28*, 124–143. https://doi.org/10.20622/jltajournal.28.0_124

Llanes, À., & Muñoz, C. (2009). A short stay abroad: Does it make a difference? *System*, *37*(3), 353–365. https://doi.org/10.1016/j.system.2009.03.001

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.

Moore, I., Torgerson, C., & Beckmann, N. (2021). Systematic review measuring the efficacy of study abroad in undergraduate language learners on linguistic proficiency gains. *Review of Education*, *9*, Article e3306. https://doi.org/10.1002/rev3.3306

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly*, *46*(4), 610–641. https://doi.org/10.1002/tesq.34

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Potential Plus K.K. (2025). *AISATS: AI Skill Assessment and Training System* [Automated speaking test]. https://potentialplus.co/ja/aisats/

R Core Team. (2025). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Saito, Y. (2025). Effects of a short-term study abroad program on students' English proficiency and foreign language anxiety. *Eigo Eibei Bungaku*, *65*, 133–153. https://chuo-u.repo.nii.ac.jp/records/2002166

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge. https://doi.org/10.4324/9780203851357

Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, *26*(2), 173–199. https://doi.org/10.1017/S0272263104262027

Sekiya, Y., Park, S., & Tsuji, R. (2018). Effects of short-term study abroad programs. *Studies in Linguistics and Language Teaching*, *29*, 161–180. https://doi.org/10.69236/0000001600

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. 239–273). John Benjamins. https://doi.org/10.1075/lllt.11.15tav

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, *29*(3), 325–344. https://doi.org/10.1177/0265532211424478

Author bio

Ken Urano is a Professor in the Faculty of Business Administration at Hokkai-Gakuen University, Japan. His research interests include second language acquisition, task-based language teaching, English for specific purposes, and computer-assisted language learning. His recent work examines the effects of short-term study abroad programs.
Email: urano@hgu.jp
ORCID: https://orcid.org/0000-0003-2550-5227