Title of Project:
Effects of Task Types on Interactional Competence in Oral Communication Assessment

Researcher:
Sonca Vo
Iowa State University
vosonca@gmail.com

Research Supervisor:
Dr. Gary Ockey
Iowa State University


Sonca Vo

Final Report

Motivation for the Research

Studies of interactional competence (IC) in oral communication assessment have highlighted problems regarding the unequal distribution of interaction patterns in interviews versus paired formats (Van Lier, 1989; Young & He, 1998). These studies, however, only looked at verbal interaction features, and no attempts in these studies were made to investigate both verbal and nonverbal interaction features elicited in interviews versus paired formats. Based on the constructivist-realist perspective, this current study investigates the theoretical construct of interactional competence. This construct is important because IC is an important sub-construct of speaking ability (Ockey & Li, 2015), and as speaking tests have evolved to measure interaction, the question of what constitutes IC is crucial in providing the valid interpretation and use of IC scores. In order to achieve the goal, the current study examines test takers' interactional performances in two contexts: interview and paired discussion to identify what interaction resources are local to interview and to paired discussion and to what extent the interview shares resources and a configuration with paired discussion.

Research Questions

1. What types of interaction features do raters attend to when rating IC in different task types?
2. What types of interaction features do the individual scripted interview tasks and paired discussion tasks elicit? To what extent does the individual scripted interview task share interaction features with the paired discussion task?
3. To what extent do interaction features contribute to variance in the IC scores across task types?

Research Methodology

Participants

There were three categories of participants: 38 test takers, six raters, and two coders. Four test takers whose video-taped performances were used to answer the first research question concerning raters' verbal reports on IC and 34 test takers whose IC performances (two performances per test taker, amounting to 68 performances) were rated holistically and analytically to answer the second and third questions with respect to the distributions of interaction features in the two task types and the variance of interaction features contributing to the IC score. The second category of participants includes six raters who were divided into two groups. The first group included four raters who produced verbal reports on test-taker performances regarding IC. The second group consisted of two raters who

provided the holistic and analytic ratings of IC for 34 test takers. The third category of participants involves two coders who coded rater verbal reports.

Data Analysis

Research Question 1 (RQ1). The four raters verbally reported on which features they attended to when judging test takers' IC. Each rater provided eight verbal reports for test takers across proficiency levels based on the EPT OC test scores: four for the individual scripted interview task and another set of four for the paired discussion task. In total, there were 32 verbal reports. After the 32 verbal reports were transcribed, the transcripts were segmented into "idea" units and then coded by two coders. An "idea" unit was defined as "a single or several utterances, either continuous or separated by other talk but falling within the same turn, with a single aspect of the performance as the focus" (Brown, Iwashita, & McNamara, 2005, p. 14). The two coders together coded four verbal reports in terms of the interaction features that raters commented in their reports.

Research Question 1 (RQ2). Sixty-eight test-taker performances were rated analytically based on the IAS. Logistic regressions were conducted to examine whether interaction features elicited in the two tasks vary. The dependent variable for logistic regressions is task type, and the independent variables are groups of interaction features. These groups of interaction features were determined using exploratory factor analysis.

Research Question 3 (RQ3). Sixty-eight test-taker performances were rated holistically based on the 4-point rating scale of IC. Hierarchical regressions were conducted separately for each task to investigate which verbal and nonverbal interaction features predict IC scores. The dependent variable is the IC score; the independent variables are groups of interaction features.

Summary of Findings

RQ1

Based on raters' verbal reports (see Figures 1 and 2), the interaction features that raters attended to in both the individual scripted interview and the paired discussion task included the following: hand gestures, body posture, eye contact, facial expressions, head nod, connecting topics, expanding topics, confirming comprehension, replying in an appropriate amount of time, initiating topics, persuading, replying with sufficient information, confirmation check, agreeing, disagreeing, self-correcting, asking questions for opinions/information, and clarification requests. Furthermore, while raters noticed *confirmation check* in the individual task, they reported on *agreeing* and *disagreeing* in the paired discussion task. Prompting elaboration, responding to requests for clarification, and correcting the interlocutor's mistakes were also the features that raters reported in the interview that they might notice more in the paired discussion task. One important aspect of the use of these features that raters commented on was related to whether these features were used sufficiently and how they were used to convey politeness. Based on these findings, an IAS was developed, which consists of five nonverbal interaction features and 17 verbal interaction features. This scale was used to seek the answers to the third and fourth research questions, which are presented in the following sections.

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

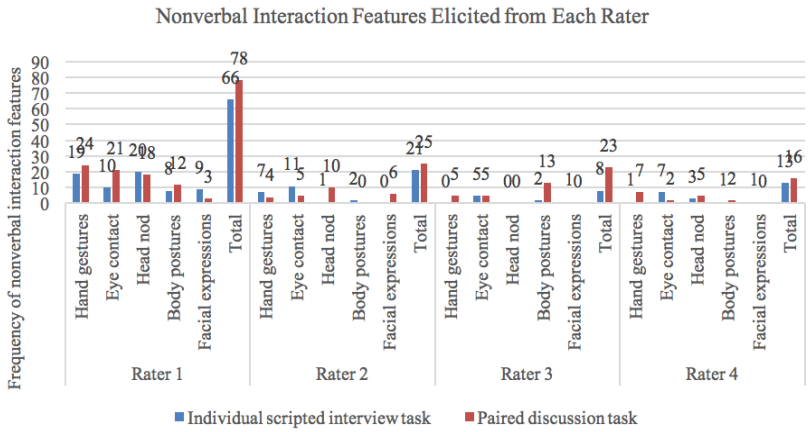

Figure 1. Nonverbal interaction features elicited from each rater

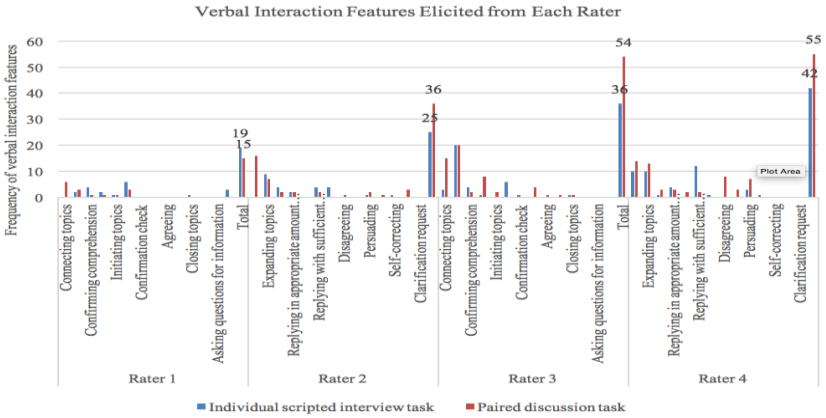

Figure 2. Verbal interaction features elicited from each rater

RQ2

Principal axis factor analysis with promax rotation was run to examine the underlying structure for the 19 features of IC. Four factors that measure body language (BL), topic management (TM), interactive listening (IL), and interactional management (IM) were extracted (see Figure 3). First, TM consists of replying with sufficient content, developing topics, replying in an appropriate amount of time, persuading, and connecting topics. Second, BL is composed of eye contact, facial expressions, hand gestures, body posture, and head nod. Third, IM consists of disagreeing, correcting mistakes made by the partner, agreeing, and comprehension check. Fourth, IL includes confirmation check, questions for opinion/information, confirming comprehension, and prompting. After rotation, the first factor accounted for 33.27% of the variance. The second factor made up 16.33% of the variance. The third factor formed 13.71% of the variance. The fourth factor constituted 6.76% of the variance.

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

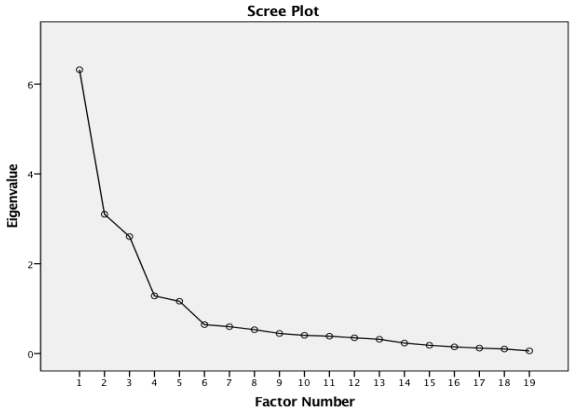

Figure 3. Scree plot

The findings from logistic regressions (see Table 1) show that the individual task elicited significantly more *connecting topics, head nod,* and *confirmation check* than the paired discussion task, with a small to large effect size. Conversely, the paired discussion task elicited significantly more *eye contact, comprehension check,* and *prompting* than in the individual scripted interview task, with a large effect size.

Predictor	*B*	*SE*	*Odds ratio*	*p*
Connecting topics (TM)	-1.09	.40	.34	.01
Head nod (BL)	-1.96	.67	.14	.00
Confirmation check (IL)	-.84	.33	.43	.01
Eye contact (BL)	1.94	.77	6.92	.01
Comprehension check (IM)	1.48	.47	4.40	.00
Prompting (IL)	1.11	.42	3.04	.01

Table 1. Logistic regression analyses

RQ3

Individual Scripted Interview Task. Multiple regression analyses with the stepwise method (see Table 2) showed that *head nod* explained the most variance to the IC scores in the individual scripted interview task. *Hand gestures* added the most next to the variance in the model, followed by *developing topics*.

Model	Predictor	*B*	*SEB*	β	*p*	R^2	R^2 *Change*
1						.56	.56
	Head nod	.87	.14	.75	.00		
2						.66	.10
	Head nod	.68	.14	.58	.00		
	Hand gestures	.31	.10	.36	.00		
3						.72	.06
	Head nod	.39	.17	.33	.03		
	Hand gestures	.27	.10	.32	.01		
	Developing topics	.31	.12	.36	.02		

Table 2. Multiple regression with stepwise method predicting IC score for the individual task

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

However, the stepwise method in multiple regression counts all of the areas where the interaction features (independent variables) overlap with IC (dependent variable) and add the interaction features one at a time to determine the best fit of the model (Tabachnick & Fidell, 2001). Using this stepwise method might have lost the unique contribution of each interaction feature to IC. Therefore, simple regression models for each interaction feature (see Table 3) were run to examine the unique contribution of each feature to the variance of IC score.

Variable	*B*	*SEB*	β	*p*	R^2
Hand gestures	.54	.12	**.63**	**.00**	.63
Body posture	.53	.14	**.57**	**.00**	.57
Eye contact	.61	.11	**.70**	**.00**	.70
Facial expressions	.62	.14	**.62**	**.00**	.62
Head nod	.87	.14	**.75**	**.00**	.75
Agreeing	-.05	.23	-.04	.84	.04
Disagreeing	-.73	.67	-.19	.28	.19
Comprehension check	.21	.21	.17	.32	.17
Confirmation check	.16	.09	.30	.08	.30
Questions for opinions/information	.06	.09	.12	.50	.12
Developing topics	.64	.10	**.75**	**.00**	.75
Connecting topics	.28	.12	**.39**	**.00**	.39
Self-correcting mistakes	.13	.12	.19	.28	.19
Correcting mistakes made by the partner	.59	.57	.18	.31	.18
Persuading	.19	.09	**.34**	**.05**	.34
Prompting	-.03	.14	-.04	.82	.04

Table 3. Simple regressions for each interaction feature

Paired Discussion Task. Hierarchical multiple regressions using the stepwise method (see Table 4) suggested that *developing topics* explained the most variance to the IC scores in the paired discussion task. *Head nod* contributed the second most variance to the scores, followed by *agreeing* which added the third most variance to the model. However, as explained in the analyses for the individual scripted interview task above, multiple regressions with the stepwise method did not consider the unique contribution of each interaction feature to the R-square. Thus, simple regression models were run to investigate the variance to the model uniquely explained by individual interaction features. The findings from simple regressions are presented in the next paragraph.

Model	Predictor	*B*	*SEB*	β	*p*	R^2	R^2 *Change*
1						.66	.66
	Developing topics	.80	.10	.82	.00		
2						.79	.13
	Developing topics	.53	.10	.54	.00		
	Head nod	.32	.07	.45	.00		
3						.83	.04
	Developing topics	.48	.10	.49	.00		
	Head nod	.26	.07	.37	.00		
	Agreeing	.13	.05	.23	.02		

Table 4. Multiple regression with stepwise method predicting IC score for the paired discussion task

Simple regressions (see Table 5) suggest that eight interaction features (i.e., hand gestures, body posture, eye contact, facial expressions, head nod, developing topics, connecting topics, and persuading) significantly predicted IC scores in the individual scripted interview task. Out of these eight features, all

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

five nonverbal features contributed most to predicting IC scores. Three verbal features contributed from moderate to high predictability to the IC score.

Variable	*B*	*SEB*	β	*p*	R^2
Hand gestures	.80	.10	.72	**.00**	.72
Body posture	.39	.10	.58	**.00**	.58
Eye contact	.62	.13	.66	**.00**	.66
Head nod	.55	.08	.78	**.00**	.78
Agreeing	.36	.08	.62	**.00**	.62
Disagreeing	.49	.11	.63	**.00**	.63
Comprehension check	.27	.11	.41	**.02**	.41
Confirmation check	.15	.10	.26	.14	.26
Confirming comprehension	.25	.08	.47	**.01**	.47
Developing topics	.80	.10	.82	**.00**	.82
Connecting topics	.56	.08	.77	**.00**	.77
Self-correcting mistakes	.33	.08	.58	**.00**	.58
Correcting mistakes made by the partner	.33	.14	.38	**.03**	.38
Replying in an appropriate amount of time	.54	.10	.69	**.00**	.69
Persuading	.53	.09	.73	**.00**	.73

Table 5. Simple regressions for each interaction feature

To summarize, the findings from simple regressions suggest that 14 interaction features, including four nonverbal features (i.e., hand gestures, body posture, eye contact, head nod), and 10 verbal features (i.e., agreeing, disagreeing, comprehension check, confirming comprehension, developing topics, connecting topics, self-correcting mistakes, correcting mistakes made by the partner, replying in an appropriate amount of time, and persuading) significantly predicted IC scores in the paired discussion task.

Implications

Theoretical Implications

This current study suggests that IC can be broken down into four sub-constructs. (1) body language (BL), (2) topic management (TM), (3) interactive listening (IL), and (4) interactional management (IM). These findings are consistent with Galaczi and Taylor (2018) in that both found three same main sub-constructs of IC: BL, TM, and IL. First, the current study suggests that BL can encompass eye contact, head nod, facial expressions, body posture, and hand gestures. Second, TM can include the following interaction features: developing topics, replying in an appropriate amount of time, persuading, connecting topics, and self-correcting mistakes. Third, this current study suggests that IL can cover confirming comprehension, confirmation check, and questions for opinion/information. The current study also found another sub-construct of IC which is interactional management. The features underlying IM can consist of agreeing, disagreeing, correcting mistakes made by the partner, and comprehension check. These findings were in line with previous research using conversation analysis (Galaczi, 2008, 2014) in that certain features of BL, IL, TM, and IM significantly explained test takers' IC performances.

Practical Implications

There are three practical implications drawn from this current study pertaining to task selection, rater training, and the development of an IC scale. First, the study provides an understanding of how task types affected the elicitation of interaction features. The findings of this study suggests that the nature

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

of interaction in different task types varies. The paired task format (e.g., paired discussion) tended to be more successful in eliciting a wider range of features of interaction than the individual task format (e.g., individual scripted interview). Although the individual scripted interview task shared a range of interaction features with the paired discussion task (e.g., hand gestures, body posture, eye contact, facial expressions, head nod, developing topics, connecting topics, persuading), it was still limited in eliciting features of natural conversation, for example, IL or IM. It cannot be denied that the interview test format is important in language testing since it provides test takers with the opportunity to demonstrate their abilities individually and since it has lower degrees of variability than the paired/group tasks (Van Moore, 2006). However, due to its limited elicitation of interactional functions, this interview test format should be used in conjunction with the paired or group format to maximize the elicitation of a wider range of interaction features. This combination of tasks makes it possible to assess the more complex aspect of IC, thus allowing for more valid inferences made based on test scores.

Second, a practical implication that can be drawn from this current study is related to rater training. Raters in this study compared test takers' performances when evaluating IC. Moreover, raters considered aspects of fluency and grammar/vocabulary in their evaluations of a test taker's IC, suggesting that their ratings show the halo effect when raters evaluate different constructs in the same way. This study suggests the necessity of providing more intensive training to raters so that they are not only focused on the individual performance but also the separate judgment for each sub-construct of a performance assessment as compared to the rating scale, not another test taker. Raters should also compare test takers to exemplary performances shown during rater training sessions, but not to other test takers that they encounter while rating. In addition, the fact that raters in this current study oriented to features that were not part of a rating rubric suggests that raters were not clear on how to assess IC. Although raters used an operational IC construct for rating, they in fact employed the theoretical IC construct by referring to many other interaction features that were not part of the IC scale. Plough et al. (2018) raised a question of whether this construct should be explicitly assessed for certain features or whether IC should be evaluated more globally. The position of this current study is that it is important to operationalize IC with certain interaction features and train raters so that they focus on only what is included in a rating rubric to improve rating reliability. This may not be a risk for validity because based on the constructivist-realist perspective to the interpretation of test taker behavior, IC can be inferred based on observed evidence from test performance which pertains to the theoretical IC construct.

Third, the nonverbal and the verbal features of interaction identified by the raters in this study can inform the development of an IC rating scale. While the appropriate use of nonverbal interaction features was mostly referred to when raters evaluated test takers' IC, this nonverbal behavior was not considered in the IC scale used in this current study. Moreover, regardless of the task types, raters attended to certain nonverbal interaction features when evaluating IC, such as hand gestures, body posture, head nod, eye contact, and facial expressions, and these features significantly affected raters' evaluations. If the nonverbal features are not captured in an IC rating scale, this important component of the IC construct may be missed, which could potentially lead to construct underrepresentation (Messick, 1989). Hence, a more thorough rating scale of IC should be developed in order to capture the complexities of this construct.

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41–71). Ottawa, ON: University of Ottawa Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26*(3), 342–366.

Brown, A., Iwashita, N., McNamara, T. F. (2005). *An examination of rater orientations and test taker performance on English for Academic Purposes speaking tasks* (TOEFL Monograph No. MS-29). Princeton, NJ: Educational Testing Service.

Cambridge Assessment English. (n.d.). Exams and tests. Retrieved from https://www.cambridgeenglish.org/exams-and-tests/qualifications/.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*, 1-47.

Chalhoub-Deville, M., & Deville, C. (2005). A look back at and forward to what language testers measure. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 815-832). Mahwah, NJ: Lawrence Erlbaum.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). New York, NY: Cambridge University Press.

Christensen, R. (1997). *Log-linear models and logistic regression* (2nd ed.). New York, NY: Springer.

Chomsky, N. (1965). *Aspect of the theory of syntax.* Cambridge, MA: MIT Press.

Fulcher, G. (2003). *Testing second language speaking*. London, UK: Pearson-Longman.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J., & Cohen, J. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: L. Erlbaum Associates.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, UK: Press Syndicate of the University of Cambridge.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL Monograph No. MS-18). Princeton, NJ: ETS.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research.* Thousand Oaks, CA: Sage Publications.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge, UK: Cambridge University Press.

Ducasse, A. M. (2014). *Interaction in paired oral proficiency assessment in Spanish.* Frankfurt/Main, Germany: Peter Lang.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26,* 423 – 443.

Eckes, T. (2015). *Introduction to Many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Bern, Switzerland: Peter Lang.

Educational Testing Service. (n.d.). About the TOEFL iBT test. Retrieved from https://www.ets.org/.

English Placement Test. (n.d.). Resources. Retrieved from https://apling.engl.iastate.edu/english-placement-test/.

Fabrigar, L. R., & Wegener, D. T. (2012). *Understanding statistics. Exploratory factor analysis.* New York, NY, US: Oxford University Press.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics: And sex and drugs and Rock "N" Roll* (4th ed.). Los Angeles, CA: Sage.

Folland, D., & Robertson, D. (1976). Towards objectivity in group oral testing. *English Language Teaching Journal, 30,* 156-167.

Fulcher, G. (2003). *Testing second language speaking*. London, UK: Pearson-Longman.

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing, 13*(1), 23-51.

Galaczi, E. D. (2014). IC across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics, 35*(5), 553 – 574.

Galaczi, E. D. (2008). Peer–peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly, 5*(2), 89–119.

Galaczi, E. D., & Taylor, L. (2018). IC: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, *15*(3), 219-236.

He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET–SET group discussion. *Language Testing*, *23*, 370–401.

He, A., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1-24). Amsterdam, The Netherlands & Philadelphia, PA: John Benjamins.

Heinze, G., & Schemper. M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine, 21*(16), 2409–2419.

Hymes, D. (1972). *Towards communicative competence*. Philadelphia, PA: Pennsylvania University Press.

International English Language Testing System (2019). IELTS Academic. Retrieved from https://www.ielts.org/.

Jacoby, S., & Ochs, E. (1995). Co-construction: An introduction. *Research on language and social interaction, 28*(3), 171-183.

Jamieson, J., Eignor, D. R., Grabe, W., & Kunnan, A. J. (2008). Frameworks for a new TOEFL. In C. A. Chapelle, J. Jamieson, & M. K. Enright (Eds.), *Building a validity for the test of English as a foreign language* (pp. 55-95). New York, NY: Routledge.

Johnson, M. (2001). *The art of non-communication: A reexamination of the validity of the Oral Proficiency Interview*. New Haven, CT: Yale University Press.

Kasper, G., & Ross, S. (2007). Multiple questions in oral proficiency interviews. *Journal of Pragmatics, 39*, 2045 – 2070.

Kasper, G., & Wagner, J. (2011). A conversation-analytic approach to second language acquisition. In D. Atkinson (Ed.), *Alternative approaches to second language acquisition* (pp. 117-142). New York, NY: Taylor & Francis.

Kline, R. B. (2005). *Methodology in the social sciences. Principles and practice of structural equation modeling* (2nd ed.). New York, NY, US: Guilford Press.

Kline, P. (1999). *The handbook of psychological testing* (2nd ed.). London, UK: Routledge.

Kline, P. (1994). *An easy guide to factor analysis*. New York, NY: Routledge.

Kramsch, C. (1986). From language proficiency to IC. *Modern Language Journal, 70*(4), 366–372.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests: A teacher's book*. Bristol, UK: Longmans, Green and Company.

Lam, D. M. K. (2018). What counts as 'responding'? Contingency on previous speaker contribution as a feature of IC. *Language Testing*, *35*(3), 377-401.

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2011). *IBM SPSS for intermediate statistics: Use and interpretation* (4th ed.). New York, NY: Taylor & Francis.

Linacre, J. M. (2017). *A user's guide to FACETS: Rasch-model computer programs (3.80.0) [computer software manual].* Beaverton, OR: Winsteps.com.

Linacre, J. M. (2010). *User's guide to Winsteps Ministep Rasch-Model computer programs*. Retrieved from http://www.winsteps.com/winman/.

May, L. (2011). IC in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*(2),127-145.

McNamara, T. (1997). Interaction in second language performance assessment: Whose performance? *Applied Linguistics, 18*(4), 446-466.

Mehan, H. (1979). *Learning lessons: Social organization in the classroom.* Cambridge, MA: Harvard University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York, NY: Macmillan Publishing Co.

Nakatsuhara, F., Inoue, C., Berry, V., Galaczi, E. D. (2017). Exploring performance across two delivery modes for the IELTS speaking test: Face-to-face and video-conferencing delivery. The IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Ockey, G. J. (2017). Approaches and challenges to assessing oral communication on Japanese entrance exams. *Japan Language Testing Association Journal, 20,* 3-14.

Ockey, G. J. (2014). The potential of the L2 group oral to elicit discourse with a mutual contingency pattern and afford equal speaking rights in an ESP context. *English for Specific Purposes*, *35*, 17-29.

Ockey, G. J. (2013). Exploratory factor analysis and structural equation modeling. In A. J. Kunnan (Ed.), *The companion to language assessment*. *Vol. III: Evaluation, Methodology, and Interdisciplinary Themes* (pp. 1224-1244). Malden, MA: John Wiley & Sons.

Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing, 26*, 161-186.

Ockey, G. J. (2006). *Making a case for the group oral discussion test: The effects of personality on the group oral's score-based inferences*. (Unpublished Ph.D. Dissertation). The University of California, Los Angeles.

Ockey, G., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing, 32*(1), 39-62.

Ockey, G. J., & Wagner, E. (2018). *Assessing L2 listening: Moving towards authenticity* (Vol. 50). John Benjamins, Philadelphia, PA.

Ockey, G. J., & Li, Z. (2015). New and not so new methods for assessing oral communication. *Language Value, 7*(1), 1-21.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System, 30*(2), 143-152.

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking test tasks. *Language Testing*, *19*, 33-56.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction*. Belmont, CA: Wadsworth.

Philp, J., Adams, R., & Iwashita, N. (2014). *Peer interaction and second language learning*. New York, NY: Routledge.

Pollitt, A., & Murray, N. L. (1996). What raters *really* pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp. 74-91). Cambridge, UK: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

Roever, C., & Kasper. G. (2018). Speaking in turns and sequences: IC as a target construct in testing speaking. *Language Testing, 35*(2), 331-355.

Rubin, D. L., & Schramm, G. (1997). The testing of L1 speaking. In C. Clapham & D. Corson. (Eds.) *Encyclopedia of language and education*, *Vol. 7: Language Testing and Assessment* (pp. 29-37). Amsterdam, The Netherlands: Kluwer Academic Publishers.

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning, 52*(1), 119–158.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Boston, MA: Pearson/Allyn & Bacon.

Taylor, L. (2000). Investigating the paired speaking test format. *University of Cambridge TESOL Examinations Research Notes, 2,* 14-15.

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23*, 489-508.

Van Moere, A. (2007). *Group oral test: How does task affect candidate performance and test score?* (Unpublished Ph.D. Dissertation). The University of Lancaster.

Wang, L. (2015). *Assessing IC in second language paired speaking tasks* (Unpublished Ph.D. Dissertation). Northern Arizona University.

Young, R. (2011). IC in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 426-443). London, UK & New York, NY: Routledge.

Young, R. (1999). Sociolinguistic approaches to SLA. *Annual Review of Applied Linguistics, 19*, 105-132.

Young, R. (1995). Conversational styles in language proficiency interviews. *Language Learning, 45*(1), 3–42.

Young, R., & He. A. (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency.* Amsterdam, The Netherlands & Philadelphia, PA: John Benjamins.

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org