Effects of L2 test-tasks on learners' speech performance

Siwon Park
Tim Murphey
Megumi Sugita

Abstract

While the use of tasks in assessing L2 learners' speaking ability has gained more attention and interest among educators, it is still not clear how those tasks influence their speech performance. In our study, we examined the effect of L2 speaking test-tasks on learners' speech performance using a series of statistical analyses. We administered two group oral, four semi-direct, and two interview tasks to 14 L2 learners of varying proficiency and examined their speech performance using the rating scores. In this paper, we report our findings with a specific focus on the extent to which the tasks differ in assessing the participants' speaking ability and if and how the two interview tasks, often employed for high-stakes decision making, differ from each other and against other test-task types we administered. In the correlational analyses, convergent as well as discriminant aspects were revealed of the test tasks and their sub-tests employed in this study. A pedagogical implication of such findings is discussed in the conclusion.

Part I

Large View Introduction to Task-Based Testing

Over the last few years we have been reading a wide range of books thanks to our grant (see Appendix 1). We have been trying to take a large view of the field and grasp where it is going and what are our foci. Murphey has also written about innovative testing happening in Asia (Murphey, 2009), and published a novel in English and Japanese criticizing the present entrance exam system in Japan (Murphey, 2010, 2011).

In our title are two key words: tasks and performance. Most language teachers and linguists are very familiar with the surge of detailed research in task-based-learning and project work the last few decades, largely following first an emphasis on interaction (Long & Porter, 1985 Allwright and Hall et al. 2011) and more recently the interactive instinct (Lee et al. 2009), indicating what Murphey calls the interaction imperative for SLA (Murphey 2011). These have been accompanied by a social turn in SLA (Block, 2003). Both of these also imply performance, in our title, which in Vygoskian research is crucial to learning (Holzman, 2009). Recording students' speech performances can not only be a method of evaluation but also a method of teaching (Murphey, 2001; Murphey and Sakaguchi, 2010). This entails allowing students to regularly watch themselves and evaluate themselves as part of the process, just as professional athletes and actors watch their performances, to see what they like and can improve. This implies not only allowing them to work within their zone of proximal development (ZPD) but to enhance their ability to adjust to different learning environments and show others how they might adjust to them, using and developing their zones of proximal adjustment (ZPA; Murphey, in press, 2013).

We highly recommend looping back to students whatever performances they have done and whatever research findings we have to better understand it from the participants perspective, letting them evaluate their own performances (Murphey

& Falout, 2010). This seems to be in line with what Deming suggested so many years ago with Total Quality Management in businesses (Thanks to Dennis Koyama for reminding us of this research, in a recent email, Sallis, 2005). What Deming was also taping into was the meaningfulness of genuine participation, and the excitement and agency (and meaningful sense of autonomy and purpose, cf. Pink, 2009) that comes with such participation. Testing need not be solely about others evaluating you, but a learning opportunity in which we gain a better grasp of who we are and where we are going.

Before we get into the details of our study, we would just like to look briefly at three or four references that may interest readers and which are influencing our views of task-based testing.

Daniel Pink's *Drive* (2009)

> "The central idea of the book is the mismatch between what science knows and what business does," (p. 145) which could easily be applied to education and other areas of our lives….While his metaphor of the computer operating system may make some humanists cringe with comparing ourselves once again to computers, it is a light-hearted analogy that actually works well, and implies that, while we influence others to a great extent, it is "we" who control the system upgrades or not. And while we can make incremental changes, we are also able to change complete systems at times. (Murphey, 2012)

Many educators still believe also that people are motivated by high scores on tests and feel punished by low scores, and that these are their main motivations. Pink's book tries to show us that this is no longer so. Thus, we need to allow our evaluation and testing systems to evolve so that they can provide more meaning than a simple pass or fail, a carrot or stick.

Robert Sternberg's *College Admissions for the 21st Century* (2010)

> What are the best criteria for university admissions decisions? This 212-page text seeks to answer that question and demonstrate why current admissions procedures are inadequate for recruiting students with not only good analytical and memory skills, but ample creativity, wisdom, and practical leadership potential as well. The author regards the typical criteria used to accept/reject college applicants – standardized test scores, high school GPAs, class rank, and course profiles – as only moderate barometers of first-year academic performance, and even less reliable as predictors of success in later life (Neufields 2011).

Thus, we need to be braver in our testing and evaluating and do more research into ways that will show us not how students perform now, but might perform in the future in terms of participation, creativity, and adaptability.

Steven Gould's *Mismeasure of Man* (1996)

> Impartiality (even if desirable) is unattainable by human beings with inevitable backgrounds, needs, beliefs, and desires. It is dangerous for a scholar even to imagine that he might attain complete neutrality, for then one stops being vigilant about personal preferences and their influences—and then one truly falls victim to the dictates of prejudice (p. 36).

This book greatly critiques The Bell Curve and its scientific racism, and the academic belief in neutrality and objectiveness. Written by a geologist, biologist, and history of science professor at Harvard, it argues greatly for social understanding of human action.

Guba and Lincoln *The Fourth Generation Evaluation* (1989)

> Differing from previously existing generations, this new approach moves evaluation to a new level, whose key dynamic is negotiation. The constructivist paradigm is espoused by the authors and shown to offer multiple advantages, including empowerment and enfranchisement of stakeholders, as well as an action orientation that defines a course to be followed. Not merely a treatise on evaluation theory, Guba and Lincoln also comprehensively describe the differences between the positivist and constructivist paradigms of research, and provide a practical plan of the steps and processes in conducting a fourth generation evaluation. (cover blurb)

This radical book proposes "full participative involvement, in which the stakeholders and others who may be drawn into the evaluation are welcomed as equal partners in every aspect of design, implementation, interpretation, and resulting action of an evaluation—that is, they are accorded a full measure of political parity and control" (p. 11).

Another area that we have been investigating and to which we see many parallels, is mirror neuron research, which basically states that we mirror our worlds and that our worlds also mirror us. So we need to be more careful about what we are able to put in front of the mirror. Here is where imagination plays a big role in creating different task types for testing. Good research should be able to make our lives better, more ecological, and fairer to the people we are working with. We are beginning to do this and we would encourage the administration to invest more in this type of research and teachers to think seriously about doing these types of research.

Part II

The Study

A large number of studies have been pursued by testing professionals with respect to how second language (L2) (test) tasks would affect learners' speech performance. A complete review and summaries of such prior studies are readily available elsewhere (e.g., Park, 2008). Therefore, in the second part of this paper, we will mainly discuss what we have done in our research with a narrower view of L2 speaking assessment using test tasks.

Research purposes and questions

The primary purpose of the current research is to examine the extent to which different EFL speaking test-tasks influence L2 learners' speech production. Among the test tasks that were included in this project, we will report in this paper only the part that compared the interview test with group oral and semi-direct test tasks to examine if and how the interview test differs from other types of speaking tests in eliciting L2 learners' speech performance.

In order to achieve the research purpose, we set our research question as follows:

- How comparable are the rating scores assigned to the examinees in performing the three test-tasks?

Method

Participants

Fourteen EFL learners at a Japanese university participated in this study. Among them six were male and the rest female students. One learner was a freshman, five sophomores, five juniors, and three seniors at the time of this research. In addition to the Japanese learners of English, two native speakers with years of English teaching and testing experiences participated in this study serving as the raters of the learners' speech performance.

Instruments

Japanese participants performed two interview tasks – patterned and structured – and also sat two group oral exams – topic discussion and information gap tasks. They also responded to four semi-direct speaking test tasks of picture description, map reading and direction giving, impromptu speech, and chart and table reading. These are semi-direct speech tasks as they were delivered using the computer, and the examinees were simply required to respond verbally to the stimuli prompted by the computer, i.e., there was no direct conversation involved.

Procedures

Table 1 below gives information as to the types of test tasks with explanations about the performance procedures and the time allocated to each of them. The group oral and the semi-direct tasks administered counter-balanced for the order of the task presentation so that we could control for the possible bias due to the presentation order. The interview test was given after the other two tests were completed, however.

Table 1Task specifications

Test-tasks	Setting	Time
Interview		
Patterned	- answer the questions predetermined by the interviewers.	7-8 min 10-12 min
Structured (summary and reading questions)	- read a passage and summarize the content, and answer the follow-up questions concerning the passage.	
Group oral		
Topic discussion	- discuss a topic with 2 or 3 other students.	12-15 min
Information gap	- complete a task by exchanging information with 2 or 3 other students.	12-15 min
Semi-direct	- complete tasks following the instructions on a computer.	20 min

The interview test and the group oral exam were rated concurrently by the two teacher raters, while the learner performance through the semi-direct test was first recorded and evaluated later using video recordings. Nonetheless, all observations were double-rated for fairer score assignments. Also, in rating speech performance and samples of different tasks, the same rubric was utilized for rating whenever possible. The interview test and the group oral exam included a unique performance category named “communicative effectiveness” and that category could not be assessed in the semi-direct test due to its non-interactive nature. For more details about how the tests were administered, refer to Park (2012, forthcoming).

Once the scores were collected and entered into EXCEL, they were first Rasch-adjusted for fair-average scores, and the average scores were used for analyses to answer the research question.

Results

The two tables – Tables 3 and 4 –report the descriptive statistics. Table 3 shows the descriptive statistics of the three test tasks, interview, group oral, and semi-direct, while Table 4 gives detailed information about each of sub-tests under the three tasks.

Table 3 Descriptive statistics of the three tasks (*N*=14)

	M	**Range**	**Min**	**Max**	***SD***
Interview	3.02	.70	2.66	3.36	.21
Group oral	3.00	1.20	2.43	3.63	.38
Semi-direct	2.91	1.31	2.24	3.55	.42

Table 4 Descriptive statistics of individual subtests under each test (*N*=14)

	M	***Median***	**Range**	**Min**	**Max**	***SD***
Interview						
Patterned	3.04	3.05	.55	2.75	3.30	.18
Summary	3.01	3.00	.94	2.69	3.63	.29
Reading questions	3.00	3.05	.85	2.55	3.40	.27
Group oral						
Discussion	3.02	2.98	1.55	2.25	3.80	.45
Information gap	2.98	3.00	1.00	2.55	3.55	.36
Semi-direct						
Picture	2.84	3.00	1.60	2.15	3.75	.51
Map	2.88	2.95	1.35	2.15	3.50	.43
Speech	2.98	3.03	1.45	2.15	3.60	.43
Chart	2.94	3.03	1.30	2.05	3.35	.43

With the test data in hand, we first compared the group means across the three test tasks to see if any difference would result between them. As the quick examination of the two tables for the descriptive statistics also reveals, no meaningful difference was observable across the test tasks and also among the sub-tests of them. We suspect, however, the small n-size may be one reason for this finding of non-significance in the mean comparisons.

Next, we ran correlational analyses and, just like what we had done with the test tasks, we checked two series of correlations – one with the three test tasks and the other with their sub-tests included. Table 5 and Table 6 report the results.

Table 5　Correlations between the three tests

	Interview	**Group oral**	**Semi-direct**
Interview	1.00	-	-
Group oral	.39	1.00	-
Semi-direct	.37	.30	1.00

Table 6　Correlations between the individual subtests under each test

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1. Interview Patterned	1.00	-	-	-	-	-	-	-
2. Interview Summary	0.49	1.00	-	-	-	-	-	-
3. Interview Reading Q	0.71*	0.66*	1.00	-	-	-	-	-
4. Group oral Discuss	0.50	0.47	0.50	1.00	-	-	-	-
5. Group oral Info gap	0.14	0.13	0.12	0.77*	1.00	-	-	-
6. Semi-direct Picture	-0.09	0.53	0.32	0.37	0.26	1.00	-	-
7. Semi-direct Map	-0.14	0.50	0.26	0.25	0.06	0.89*	1.00	-
8. Semi-direct Speech	0.11	0.54	0.32	0.51	0.22	0.82*	0.92*	1.00
9. Semi-direct Chart	-0.01	0.49	0.29	0.28	-0.01	0.76*	0.75*	0.78*

* significant at p < .05

As shown in Table 5, the three test tasks were found unrelated with each other in assessing the Japanese learners' English speech in this study. The marginal size of correlation coefficients indicates that basically the three test tasks are tapping different aspects of speech, although they are all measures of L2 speaking.

Yet, Table 6 informs us of a different aspect of the subtest tasks under each category. As highlighted with thick lines, except for the one between Interview Patterned and Interview Summary, all the other correlation coefficient resulted with statistical significance and their sizes are not negligible mostly over .70. These significant coefficients that fall under the same test task indicate that the sub-tests all are measuring the similar or the same construct of L2 speaking, a convergent aspect of the measurement.

Discussions and Conclusion

We performed a series of analyses with the rating scores. While the mean comparisons did not produce any statistically meaningful differences, follow-up correlations revealed several interesting findings about the test tasks in relation to each other in this research. First, the reading question section of the interview test was highly correlated with the other two sub-tests of the interview test; yet, it was not related to other tests under the different types of tasks. Second, two group oral tasks were highly correlated with each other. Third, four semi-direct tasks were all highly correlated with each other, e.g., basically, the four subtests, Picture, Map, and Speech test tasks appear to be tapping the same speaking ability trait.

The findings through the correlational analyses present empirical evidence for discriminant validity of the three tasks. The subtests of the same method are essentially measuring the same aspect of the speaking ability by L2 learners in this study. One implication of such discriminant aspect of the test tasks is the needs for diversifying the methods of testing as well as teaching. One may easily fail measuring L2 learners' speaking ability only by employing a single method. Furthermore, one's teaching of L2 speaking may not be complete if she/he only utilizes a single teaching method. As we argued earlier, if we acknowledge the importance of putting teaching and testing along the same instructional cycle, we must employ a variety of methods of testing and teaching so that pedagogically sound assessment and the feedback from it could be provided to L2 learners .

Before we conclude our study, we'd like to note a couple of things regarding our research. First, we were not able to find any meaningful group differences between the three test tasks. We suspect such statistical inability is due to our small sample size, which is the reason for our low statistical power. Inviting a few more participants and adding their rating scores to the data may have helped statistically differentiate the tasks. Although the possibility is weak, the convergent-discriminant aspect of the data revealed by the correlational analyses might be simple artifacts due solely to rating dependency. However, considering that the two raters who served in the research were experienced teachers and testers with extensive training in rating, we assume it unlikely that they were simply responding to the testing method inflating the method effect in their rating process. Finally, in this paper, we are not reporting our second empirical study of speech sample analyses. Examining the speech samples produced by the participants for their linguistic sophistication would have helped us achieve the research purpose better – the effect of L2 tasks on learner speech performance. We leave the speech sample analyses to a future paper.

Finally, we wish to call once again on other researchers to creatively examine the effects of various L2 test-tasks on learners' speech performances, and for teachers to study the effects of a variety of tasks on learners' speech performances in the classroom itself. We once again wish to acknowledge the importance of putting teaching and testing along the same instructional cycle, employing a variety of methods of testing and teaching so that pedagogically sound and diverse performance assessments and their feedback can be provided to L2 learners. When students have clear test feedback and are knowledgeable about pedagogical tasks to improve performances, they can feel more in control and motivation increases.

Appendix 1 References

Murphey's Primary Sources

Johnson, S. (2010). *Where Good Ideas Come From: The Natural History of Innovation.* New York: Riverhead Books.

Pink, D. (2009). *Drive: The Surprising Truth About What Motivates Us*. NY: Penguin Sternberg, R. (2010). *College Admissions for the 21st Century*. Cambridge: Harvard University Press.

Secondary Sources

Johnson, S. (2001). Emergence: *The connected Lives of Ants, Brains, Cities, and Software*. NY: Scribner.

Quinn, J. (2010). *Learning Communities and Imagined Social Capital: Learning to belong*. London: Continuum.

Rifkin, J. (2009). *The Empathetic Civilization: The Race to Global Consciousness in a World in Crisis.* NY: Penguin.

de Waal, F. (2009).*The Age of Empathy: Nature's Lessons for a Kinder Society.* NY: Harmony Books.

Outside of Grant Money

Holzman, L. (2009). *Vygotsky at Work and Play*. London: Routledge.

Gould, S. (1981, 1996 revised). *The Mismeasure of Man*. New York: W.W. Norton & Company. (Well written. Exposes bad use of IQ testing. Good Chapter on Factor Analysis)

Guba, E. and Lincoln Y. (1989). *Fourth Generation Evaluation*. London: Sage. (Negotiated evaluation, empowerment and enfranchisement of stakeholders. Post-positivism. Parallels Deming's TQM)

Sacks, P. (1999). *Standarized Minds: The high price of America's testing culture and what we can do to change it.* New York: Da Capo Press.

Perez, E. (March 24, 2008). No-test option gives Lawrence a different look. *Journal Sentinel, Inc*. Accessed July 6, 2012 at http://www.jsonline.com/news/education/29534914.html

Bios:

Siwon Park teaches in the English Department at Kanda University of International Studies. He has a PhD from the University of Hawaii at Manoa, and his research interests include second language acquisition, psycholinguistics, language testing, and statistics.

Tim Murphey (PhD Université de Neuchâtel, Switzerland) researches Vygotskian socio cultural theory (SCT) with a transdisciplinary emphasis on community, play, and music at Kanda University of International Studies.

Megumi Sugita teaches at Chiba Prefectural University of Health Sciences. She studied in the M.A. program both at Yokohama National University and at the University of Hawaii at Manoa. Her research interest includes motivation in second language learning, cross-cultural understanding in language teaching, and second language learners' identity.

References For Part I

Allwright, D., & Hanks, J. (2009). *The developing language learner*. New York: Palgrave Macmillan.

Block, D. (2003). *The social turn in second language acquisition*. Washington, D.C.: Georgetown University Press.

Lee, N., Dina, A. Joaquin, A., Mates, A., & Schumann, J. (2010). *The Interactional Instinct. New York*: Oxford University Press.

Long, M., & Porter, P. (1985). Group work, interlanguage talk, and second language acquisition. *TESOL Quarterly, 19*(2) 206-28

Murphey, T. (2001). Videoing conversations for self evaluation in Japan. In J. Murphy & P. Byrd (Eds.) *Understanding the courses we teach: Local perspectives on English language teaching*. pp. 179-196. Ann Arbor: University of Michigan Press.

Murphey, T. (2009). Innovative School-Based Oral Testing in Asia. *Shiken: JALT Testing & Evaluation SIG Newsletter* Vol. 13 No. 1. January 2009. (p. 14 - 21) Accessed at http://jalt.org/test/mur_4.htm

Murphey, T. (2010). *The Tale that Wags*. Nagoya, Japan: Perceptia Press.

Murphey, T. (2011). ゆれ動くしっぽ Ure ugoku Shipo. (Tale that Wags, in Japanese). Nagoya, Japan: Perceptia Press.

Murphey, T. (2011). The L2 Passionate Interactional Imperative (for short "The L2 Pie"): It's hot or it's not! *Studies in Self-Access Learning Journal, 2*(2), 87-90.

Murphey, T. (in press 2013). Adapting ways for meaningful action: ZPDs and ZPAs. In J. Arnold & T. Murphey (Eds.), *Meaningful action: Earl Stevick's influence on language teaching. Cambridge*: Cambridge University Press.

Murphey, T., & Falout, J. (2010). Critical participatory looping: Dialogic member checking with whole classes. *TESOL Quarterly, 44*(4) 811-821.

Murphey, T. & Sakaguchi, J. (2010). Multitasked student video recording. In A. Ahhadeh & C. Coombe (eds.) *Applications of task-based learning in TESOL.* Ch. 8, pp. 97-110. Alexandria Va: TESOL

Neufields, T. (2011). Book Review: *College Admissions for the 21St Century.* Robert Sternberg. Cambridge: Harvard University Press (2010).

Sallis, E. (1993/1996/2002/2005). *Total Quality Management in Education. London:* Taylor and Francis e-Library

References For Part 2

Park, S. (2008). Differential effects of EFL tasks on learners' speech production. *Studies in Linguistics and Language Teaching, 19*, 185-205.

Park, S. (to appear 2012). Comparability of tasks in assessing L2 learners' speaking performance. *The Journal of Kanda University of International Studies, 25.*