Examining Generalizability of Speaking Performance Measures Across Tasks

Rie Koizumi (Tokiwa University, Japan
e-mail: rie-koizumi@mwa.biglobe.ne.jp)
1. Introduction

Generalizability:

“The extent to which research results can justifiably be applied to a situation beyond the research setting” (Chalhoub-Deville, 2006, p. 3)

“Language tests are procedures for generalizing” (McNamara, 2006) and we always need to generalize to some degree.

Generalizability is an essential aspect of validity (e.g., Kane, 2001; Messick, 1996)

BUT it is problematic in assessing language, including speaking (Fulcher, 1997).
Chalhoub-Deville (2003): “The need for a theory of context” (p. 281)
We need to know in what aspect and to what degree we can generalize results of speaking performance (measures) across what type of contexts.

Previous studies: Lee (2006), Weir & Wu (2006)

They report some degree of generalizability of speaking when rating scales are used.

→What about generalizability when speaking performance measures are used?

Purpose of this study:

To investigate the generalizability of speaking performance across tasks when speaking performance measures are used.
Speaking performance measures: “discourse analytic measures,” which “provide counts of specific linguistic features occurring in the discourse” (Ellis, 2003)

Substantially utilized in SLA research (e.g., Skehan & Foster, 2001) without much examination of their qualities

Example: “No. of error-free clauses” divided by “No. of clauses” for accuracy

Research question 1: What aspect of speaking performance can be more generalizable?

Research question 2: Across what type of tasks can speaking performance be more generalizable?
Significance of this study:
This study can help researchers and teachers make more valid inferences, select appropriate measures, as well as understand the importance of using multiple measures and tasks.

2. Method (Koizumi, 2005)
Participants: 225 Japanese EFL learners at the novice level
Speaking Test: Tape-mediated with 5 monologue tasks
This study only analyzed 4 tasks (see Table 1).
Tasks 2 & 5: Same structure (i.e., comparison of two pictures) with different difficulty (Task 5 is more difficult than Task 2)
Tasks 3 & 4: Same structure (i.e., description of one picture) with similar difficulty

<table>
<thead>
<tr>
<th>Task</th>
<th>Speaking type</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Description (Not used in this study)</td>
<td>Self-introduction</td>
</tr>
<tr>
<td>2</td>
<td>Picture comparison</td>
<td>Comparing pictures on the left and the right</td>
</tr>
<tr>
<td>3</td>
<td>Picture description</td>
<td>Washing dishes</td>
</tr>
<tr>
<td>4</td>
<td>Picture description</td>
<td>Riding bicycles</td>
</tr>
<tr>
<td>5</td>
<td>Picture comparison</td>
<td>Comparing Taro’s rooms before and after</td>
</tr>
</tbody>
</table>
Analyses
Assessed from 4 aspects: (a) Fluency, (b) accuracy, (c) syntactic complexity, and (d) lexical complexity (see Table 2 for the measures used)

Analysed from 2 perspectives
(x) Correlations between Tasks 2 & 5 and between Tasks 3 & 4, using $r$
(y) Differences in means between Tasks 2 & 5 and between Tasks 3 & 4, using Hedges’s $g$ (effect size)

Interpretation (Cohen, 1988):
Lower than small $< .20; .20 \leq$ small $< .50; .50 \leq$ medium $< .80; .80 \leq$ large
Table 2  Summary of Correlations and Effect Sizes for Differences in Means

<table>
<thead>
<tr>
<th>Construct</th>
<th>Code</th>
<th>Measure of Study 2D</th>
<th>T2 &amp; T5 r (95%CI)</th>
<th>T3 &amp; T4 r (95%CI)</th>
<th>T2 &amp; T5 g (95%CI)</th>
<th>Interpretation</th>
<th>T3 &amp; T4 g (95%CI)</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>F1</td>
<td>No. of tokens per minute</td>
<td>.71 (.64 to .77)</td>
<td>.74 (.67 to .79)</td>
<td>-.03 (-.13 to .07)</td>
<td>zero</td>
<td>-.24 (-.34 to -.15)</td>
<td>zero, small</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>No. of dysfluency markers per minute</td>
<td>.50 (.40 to .59)</td>
<td>.57 (.47 to .65)</td>
<td>-.05 (-.18 to .08)</td>
<td>zero</td>
<td>.15 (.03 to .27)</td>
<td>zero, small</td>
</tr>
<tr>
<td>Accuracy</td>
<td>A1</td>
<td>No. of error-free clauses per clause</td>
<td>.01 (-.12 to .14)</td>
<td>.32 (.20 to .43)</td>
<td>-.29 (-.47 to -.10)</td>
<td>zero, small</td>
<td>.17 (.01 to .32)</td>
<td>zero, small</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>No. of error-free AS-units per AS-unit</td>
<td>.04 (-.09 to .17)</td>
<td>.31 (.19 to .42)</td>
<td>-.32 (-.50 to -.14)</td>
<td>zero, small, medium</td>
<td>.11 (-.05 to .26)</td>
<td>zero, small</td>
</tr>
<tr>
<td>Syntactic complexity</td>
<td>SC1</td>
<td>No. of clauses per AS-unit</td>
<td>.19 (.06 to .31)</td>
<td>.17 (.04 to .29)</td>
<td>-.17 (-.34 to -.00)</td>
<td>zero, small</td>
<td>.48 (.30 to .65)</td>
<td>small, medium</td>
</tr>
<tr>
<td></td>
<td>SC2</td>
<td>No. of tokens per AS-unit</td>
<td>.31 (.19 to .42)</td>
<td>.35 (.23 to .46)</td>
<td>-.44 (-.60 to -.28)</td>
<td>small, medium</td>
<td>.30 (.15 to .45)</td>
<td>zero, small</td>
</tr>
<tr>
<td>Lexical complexity</td>
<td>LC1</td>
<td>Guiraud index: No. of types divided by the square root of No. of tokens</td>
<td>.26 (.13 to .38)</td>
<td>.56 (.46 to .64)</td>
<td>-.88 (-1.06 to -.70)</td>
<td>medium, large</td>
<td>-.32 (-.45 to -.20)</td>
<td>zero, small</td>
</tr>
<tr>
<td></td>
<td>LC2</td>
<td>No. of lexical word types per token</td>
<td>.32 (.20 to .43)</td>
<td>.34 (.22 to .45)</td>
<td>.08 (-.07 to .24)</td>
<td>zero, small</td>
<td>.02 (-.13 to .17)</td>
<td>zero</td>
</tr>
<tr>
<td></td>
<td>LC3</td>
<td>No. of sophisticated word types per token using J101</td>
<td>.40 (-.03 to .23)</td>
<td>.32 (.20 to .43)</td>
<td>.40 (.26 to .55)</td>
<td>small, medium</td>
<td>.19 (.04 to .34)</td>
<td>zero, small</td>
</tr>
</tbody>
</table>

Note. J1001 = The criterion of considering 101 to 8,000 words in the JACET8000 (JACET, 2003) as sophisticated.
3. Results

(x) Correlations between Tasks 2 & 5 and between Tasks 3 & 4, using $r$

Research question 1: What aspect of speaking performance can be more
generalizable?

Speaking speed (F1) > Percentage of dysfluency markers (F2), Lexical
diversity (LC1; when tasks are similar in structure and difficulty) > the rest

Research question 2: Across what type of tasks can speaking performance be
more generalizable?

More generalizable when tasks are similar in structure and difficulty than
when tasks are similar in structure but different in difficulty
(y) Differences in means between Tasks 2 & 5 and between Tasks 3 & 4, using Hedges’s $g$ (effect size)

Research question 1: What aspect of speaking performance can be more generalizable?

Tasks 2 & 5: F1, F2, A1, SC1, LC2 are more generalizable.

Tasks 3 & 4: F1, F2, A1, A2, SC2, LC1, LC2, LC3 are more generalizable.

Research question 2: Across what type of tasks can speaking performance be more generalizable?

More generalizable when tasks are similar in structure and difficulty than when tasks are similar in structure but different in difficulty.

When tasks are similar in structure and difficulty, all aspects had zero or small effect sizes but SC1.
4. Conclusion

Research question 1: What aspect of speaking performance can be more generalizable?

→ The generalizability depends on the measures used, but generally speaking, fluency (especially speaking speed, as measured by F1) is more generalizable than other aspects.

→ When a construct is fluency (especially speaking speed), the trait position, rather than the interactionalist position (Chapelle, 1998) can be taken in defining and interpreting constructs.
Research question 2: Across what type of tasks can speaking performance be more generalizable?

→ There are task differences in correlations and means to some degree, but speaking performance can be more generalizable when tasks are similar in structure and difficulty than when tasks are similar in structure but different in difficulty.

Further research: Generalizability should be examined across more tasks and other settings.

References


Chalhoub-Deville, M. (2006). Drawing the line: The generalizability and limitations of research in applied linguistics. In M. Chalhoub-Deville, C. A. Chapelle, & P. Duff (Eds.), Inference and generalizability in applied linguistics: Multiple


Test setting:
An assistant language teacher (Ms. Smith) is a new teacher. She wants to know your English ability, so you will have an English speaking test. Speak as if you were speaking to her. You do not have time to prepare. When Ms. Smith says “Please start” please say “Yes” and start speaking right away.

Task 2
There are differences between the two pictures. Please explain the differences. Please talk about the marked objects first. If you do not know what to say, please talk about anything. You have 60 seconds to speak. When Ms. Smith says “Please start” please say “Yes” and start speaking right away.

Task 3
Describe this picture in as much detail as possible so that Ms. Smith, who is not looking at the picture, can understand what is in it. Please talk about the marked behaviors first. If you do not know what to say, please talk about anything. You have 60 seconds to speak. When Ms. Smith says “Please start” please say “Yes” and start speaking right away.

Task 4
Describe this picture in as much detail as possible so that Ms. Smith, who is not looking at the picture, can understand what is in it. Please talk about the marked behaviors first. If you do not know what to say, please talk about anything. You have 60 seconds to speak. When Ms. Smith says “Please start” please say “Yes” and start speaking right away.

Task 5
Please unfold the folded picture. There are pictures above and below. Your brother (Jiro) is mischievous. While you were away at school, he scattered your belongings in your room. Say how and what in the room has changed by saying “something was something before, but now something is something.” If you do not know what to say, please talk about anything. You have 60 seconds to speak. When Ms. Smith says “Please start” please say “Yes” and start speaking right away.