Reports of 2011 Studies in Japan Association for Language Education and Technology, Kansai Chapter, Methodology Special Interest Groups (SIG)
Yo In'nami & Rie Koizumi (pp. 15–40)

Reproduction of Structural Equation Models in Second Language Testing and Learning Research

Yo In'nami
Shibaura Institute of Technology
Rie Koizumi
Tokiwa University

Abstract

Despite the prevalent use of structural equation modeling (SEM) in second language testing and learning and the growing awareness of the value of replication studies as a means for knowledge accumulation and critical appraisal of previous studies, the existing structural equation models have been infrequently reproduced, with the exception of Fulcher (1996). This article summarizes the significance of the reproduction of structural equation models in second language testing and learning, illustrates the procedures for such reproduction, and discusses the value and difficulties of such reproduction by answering frequently asked questions.

Keywords: reproduction/replication/reanalysis, structural equation modeling, variance/covariance matrix, correlation matrix

1. Introduction

Recently, an increasing amount of second language learning literature has emerged regarding the importance of replication studies (Abbuhl, 2012; Gass & Mackey, 2007; Language Teaching Review Panel, 2008; Mackey & Gass, 2005; Polio & Gass, 1997; Porte, 2009, 2010; Santos, 1989; Valdman, 1993, 1997), in which previous studies are systematically manipulated in order to examine whether earlier findings can be replicated. The long-time paucity of replication studies in second language testing and learning, as well as in the social sciences, is owing to (a) editorial and academic preferences for novelty, (b) the tendency to view successfully replicated findings as trivial, and (c) the misinterpretation of statistical *p* values as the probability of replication (e.g., Kline, 2004; Language Teaching Review Panel, 2008). However,

replication has been gaining prominence for two primary reasons. First, replication is a useful procedure for critically appraising empirical and theoretical findings from previous studies. The successful replication of previous findings—and, thereby, the confirmation of their veracity—ensures that researchers can safely base new studies on earlier work (e.g., Hedges, 1987; Kline, 2004). Second, replication prompts researchers to take a more critical stance on the findings of previous studies, because such an understanding of the original studies is an essential prerequisite for good replication studies (King, 1995; Santos, 1989).

Replication can be classified into four types: exact, approximate, constructive, and internal replication (Finifter, 1975; Language Teaching Review Panel, 2008; La Sorte, 1972; Lykken, 1968). One type of replication that can also be classified into exact replication—and on which there has not been much focus—is model replication in the context of structural equation modeling (SEM). Since SEM has been widely used in second language testing and learning, the insights we can gain into replication (or more conventionally referred to as reproduction in the SEM literature) of studies using SEM would contribute greatly to the growing awareness of the value of replication. This paper discusses the significance of the reproduction of structural equation models in second language testing and learning, illustrates the procedures for such reproduction (see also In'nami, 2011), and discusses the value and difficulties of such reproduction by responding to frequently asked questions.

2. Literature Review

Until recently, replication had not gained much attention in second language learning. There are three main reasons for this. First, replication studies are often associated with a lack of innovativeness or originality and are thereby believed to hold little appeal to a large readership (e.g., Mackey & Gass, 2005; Ortega, 2009; Porte, 2009, 2010; Santos, 1989). Second, obtaining consistent results across the original and replicated study is considered trivial (thus contributing little to the literature; e.g., Gass & Mackey, 2007; Language Teaching Review Panel, 2008). Third, *p* values are misunderstood as the probability that a result will be replicated (thereby suggesting that there is no need to conduct replication studies), which is shown to be a misconception (Carver, 1978; Kline, 2004). However, the Language Teaching Review Panel (2008) argued that "replication studies can verify and consolidate previous findings, helping results to be converged and extended. Thus, they must be more valued, encouraged and carried out in our field" (p. 11). Echoing the need for more

replication studies, Polio and Gass (1997) stated that "some of the so-called facts of our field have been determined on the basis of scanty evidence" (p. 500). Similarly, Valdman (1993), the editor of the journal *Studies in Second Language Acquisition*, argued that "the way to more valid and reliable SLA research is through replication" (p. 505), and since then, the journal has devoted a special section to replication studies. In line with this emphasis on replication studies, the journal *Language Teaching* invites the submission of articles based on replication studies.

It should be recognized that replication is useful in two ways. First, irrespective of whether replication is successful, it helps accumulate and advance knowledge in a particular field. According to Hedges (1987) and Kline (2004), successful replication is a requirement of scientific inquiry, allows stronger claims to be made about the verification of the original study, and suggests the generalizability of study findings for different contexts; unsuccessful replication raises questions about the value of the original study and leads to a reconsideration or modification of the original findings. Similarly, Titscher, Meyer, Wodak, and Vetter (2000) argued that before any scientific work is accepted, it needs to be verifiable, replicable, and repeatable. Second, replication provides students and scholars with an opportunity to read the existing literature carefully and critically and to understand the exact process by which the results were produced (Fitzpatrick, 2009; Santos, 1989). Such a process, according to King (1995), is "*the only way to understand and evaluate an empirical analysis fully*" (italics in original, p. 444).

Literature on replication distinguishes between four types of replication, depending on the closeness between the original and replicated studies. First, exact (or literal) replication is conducted in a way in which the original study is duplicated as exactly as possible, including subject populations, sampling methods, design, procedures, outcome measures, and data analysis (Language Teaching Review Panel, 2008; Lykken, 1968). Second, approximate (or systematic or operational) replication consists of an exact duplication of some of the essential variables (e.g., experimental procedures) of the original study (e.g., Lykken, 1968). Approximate replication seeks to examine whether one can duplicate a result using the same methods reported in the original study. The results will show how dependent a finding is on the particular research conditions. Third, constructive (or conceptual) replication entails a drastic modification of the original design (e.g., Lykken, 1968). A researcher conducts the new study by redesigning all aspects of it. A successful constructive replication verifies the result of the original study, and an unsuccessful constructive replication

needs to be carefully interpreted to identify what aspect of the new study was responsible for the divergent results. The nature of the phenomenon in question, for example, may change depending on how it is operationalized or measured. Fourth, internal replication (La Sorte, 1972), also called pseudoreplication (Finifter, 1975), involves collecting data for both the original and replication studies at the same time. Such simultaneous collection of data allows one to cross-validate the results and interpret divergent results between the original and replication studies as being uninfluenced by the lapse of time.

Among the four classifications of replication, exact replication is often seen as unrealistic because of the inherent difficulty in conducting a new study without changing the nature of the variables from the original study (Language Teaching Review Panel, 2008). We argue, however, that exact replication could also include reanalysis of the data from the original study to verify that the analysis had been correct (i.e., reproduced), using SEM. Replication usually means collecting the new data and investigating whether the results using different samples yield the same outcomes as the original research. However, the use of the term "replication" to mean a "reanalysis" of the original data to examine if the same findings are obtained is seen in Fienberg, Martin, and Straf (1985), Freese (2007), Hulland, Chow, and Lam (1996), and King (1995). Thus, reanalysis or reproduction can be considered as replication and classified in the category of exact replication. The three terms "reanalysis," "reproduction," and "replication" are used interchangeably throughout this paper.

SEM is a statistical method used to investigate the nature of relationships among variables by adopting a confirmatory, hypothesis-testing approach to the data (e.g., Bollen, 1989). SEM has been used in second language testing and learning research to examine, for instance, the factor structure of tests or questionnaires to investigate abilities or traits assessed and their interrelationships (e.g., Gorsuch, 2000; Lee, 2005). Using SEM, one can reanalyze the original model(s) reported in an article if the necessary information for reanalysis is available (see Byrne, 2006; In'nami & Koizumi, 2011; Kline, 2011 for SEM).

In order to replicate or reproduce the structural equation models, we need a (a) variance/covariance matrix or (b) correlation matrix (preferably along with standard deviations [*SD*s]). Reproduction with a variance/covariance or correlation matrix generally produces similar results if the original models were correctly reported and if the reproduction was performed correctly. We will demonstrate the procedures in the next section.

Together with the general value of replication delineated above (i.e., enhancement of knowledge accumulation and an in-depth understanding of the original studies), reproduction using SEM for exact replication has two other merits. First, reproduction helps identify specific problems in the original study, such as the accuracy of reporting values, unless the same or very similar result is reproduced, which enables researchers to take a more critical stance. Second, through reproduction, researchers can test competing (i.e., alternative, rival) models that the original authors did not test. This would facilitate the comparison of competing models and contribute to the building of theories, which makes the research more fruitful and accumulative.

Although the reproduction of previous structural equation models by using their variance/covariance or correlation matrices has been common in books and articles on SEM (e.g., Bollen, 1989; Brown, 2006; Raykov & Marcoulides, 2006), few attempts at reanalysis have been made in the field of second language testing and learning. Although they did not exactly reanalyze previous structural equation models, similar attempts are seen in Fulcher (1996), which reanalyzed Dandonoli and Henning's (1992) multitrait-multimethod correlation matrix of the American Council on the Teaching of Foreign Languages (ACTFL) rating scales in the framework of SEM. Fulcher (1996) reported a poor fit of the model with the data, suggesting the insufficiency for a validity claim to be made as opposed to the authors' validity argument for the scales. It should be noted that Fulcher (1996) did not exactly conduct a reproduction of previous structural equation models, because Dandonoli and Henning (1992) interpreted the correlation matrix and did not use SEM and because Fulcher reanalyzed their matrix by using SEM. In'nami and Koizumi (2010) systematically attempted to reproduce SEM models in the second language testing and learning field, which is briefly described in section 4.

3. Procedure for Replicating Structural Equation Models

This section illustrates the reproduction of SEM models using (a) a variance/covariance matrix and (b) a correlation matrix. Since correlation matrices are usually more often reported than variance/covariance matrices, our illustration is primarily based on correlation matrices. Nevertheless, the procedures for reproduction are the same regardless of whether a variance/covariance matrix or a correlation matrix is the input for the SEM model. The first correlation matrix we use is from Choi, Kim, and Boo (2003), which investigated the comparability of a paper-based and a

computer-based language test using confirmatory factor analysis. The model they reported is shown in Figure 1.

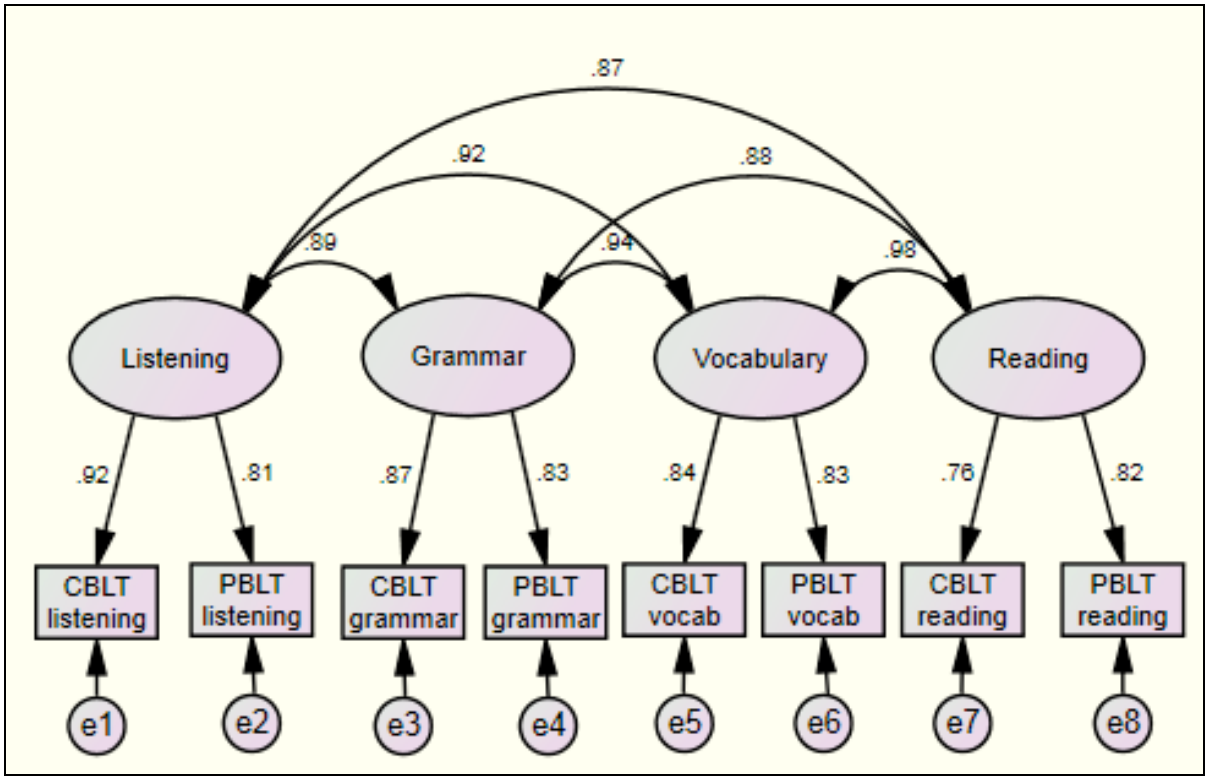
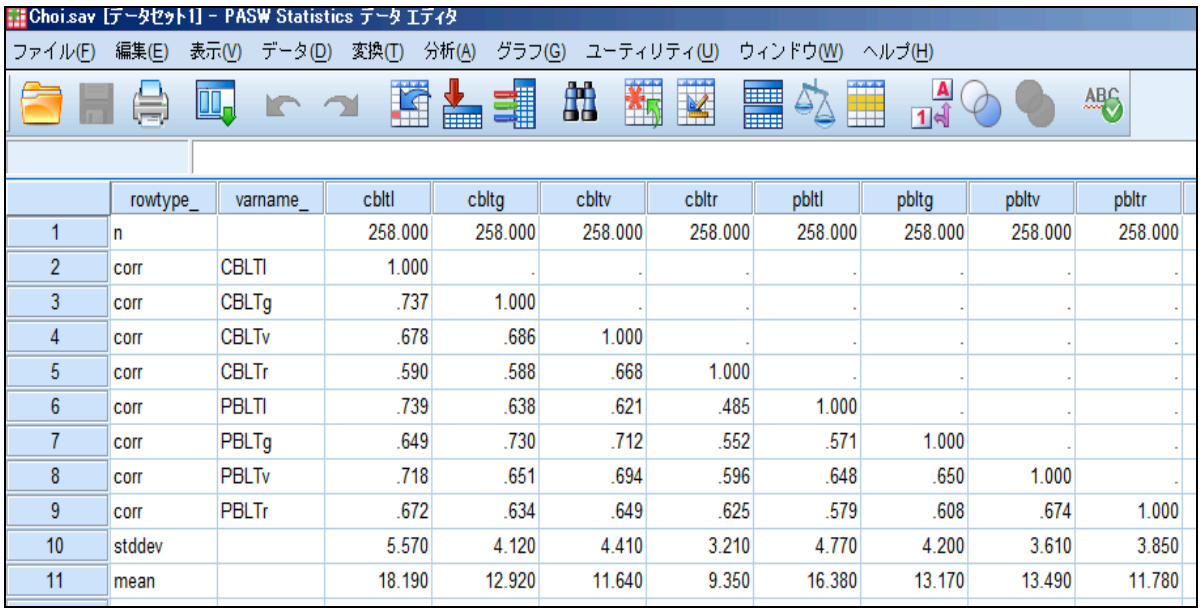


Figure 1. A model of four abilities measured across two test batteries. Adapted from Choi et al. (2003, p. 313). CBLT and PBLT refer to computer-based and paper-based language tests, respectively. All parameter estimates are standardized.

The input data were a correlation matrix and *SD*s (see Tables 9 and 13 in Choi et al., 2003) and should be entered into SPSS or Microsoft Excel, as in Figure 2. Note that the means are not used in the current reproduction but become necessary for reproduction of models such as latent growth models (e.g., Kline, 2011).



Choi.sav [データセット1] - PASW Statistics データ エディタ

ファイル(F) 編集(E) 表示(V) データ(D) 変換(T) 分析(A) グラフ(G) ユーティリティ(U) ウィンドウ(W) ヘルプ(H)

	rowtype_	varname_	cbltl	cbltg	cbltv	cbltr	pbltl	pbltg	pbltv	pbltr
1	n		258.000	258.000	258.000	258.000	258.000	258.000	258.000	258.000
2	corr	CBLTl	1.000	.	.	.	.	.	.	.
3	corr	CBLTg	.737	1.000	.	.	.	.	.	.
4	corr	CBLTv	.678	.686	1.000	.	.	.	.	.
5	corr	CBLTr	.590	.588	.668	1.000	.	.	.	.
6	corr	PBLTl	.739	.638	.621	.485	1.000	.	.	.
7	corr	PBLTg	.649	.730	.712	.552	.571	1.000	.	.
8	corr	PBLTv	.718	.651	.694	.596	.648	.650	1.000	.
9	corr	PBLTr	.672	.634	.649	.625	.579	.608	.674	1.000
10	stddev		5.570	4.120	4.410	3.210	4.770	4.200	3.610	3.850
11	mean		18.190	12.920	11.640	9.350	16.380	13.170	13.490	11.780

Figure 2. SPSS input with a correlation matrix, means, and *SD*s. rowtype_ = row type. varname_ = variable name. cbltl = CBLT Listening. n = sample size. corr = correlation. stddev = *SD*.

Subsequently, we specify and read the input file on Amos (Arbuckle, 1994–2011), as in Figure 3, which used Amos (Version, 18.0.0; Arbuckle, 2009). Reproduction can also be conducted using other SEM software programs such as EQS (Bentler, 1994–2011) and Mplus (Muthén & Muthén, 1998–2011).

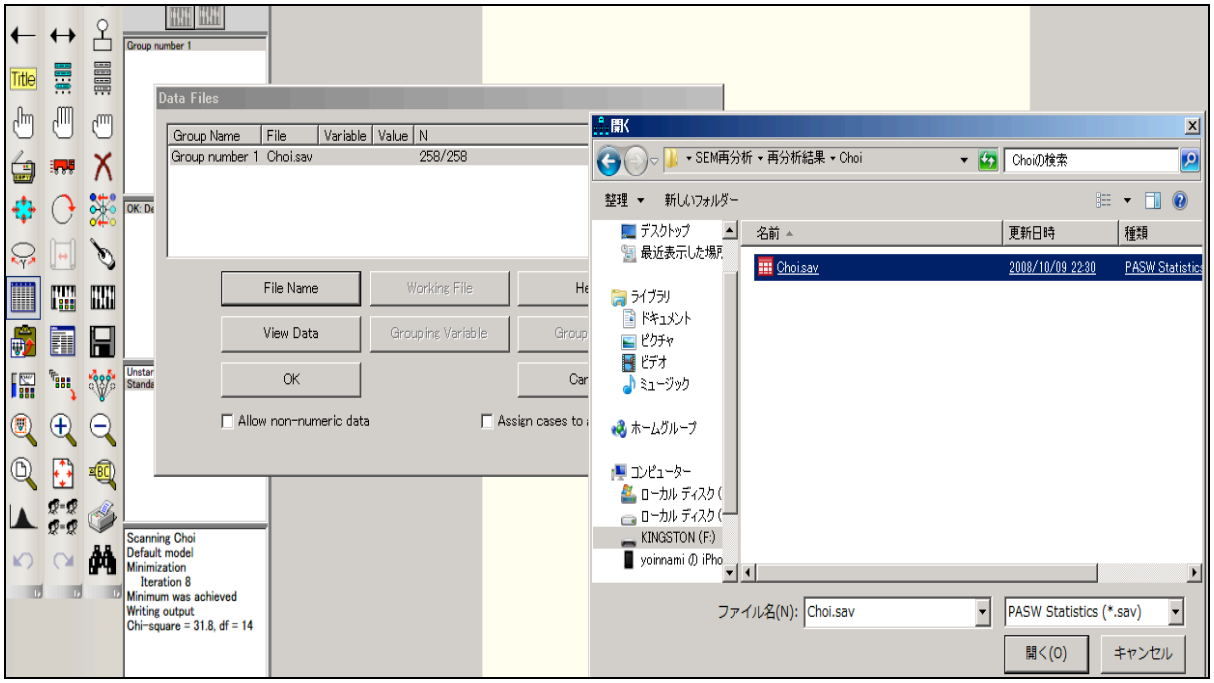


Figure 3. Input file specification.

We construct the model as shown in Figure 4 above, and then run the model. The reproduced model is presented in Figure 5.

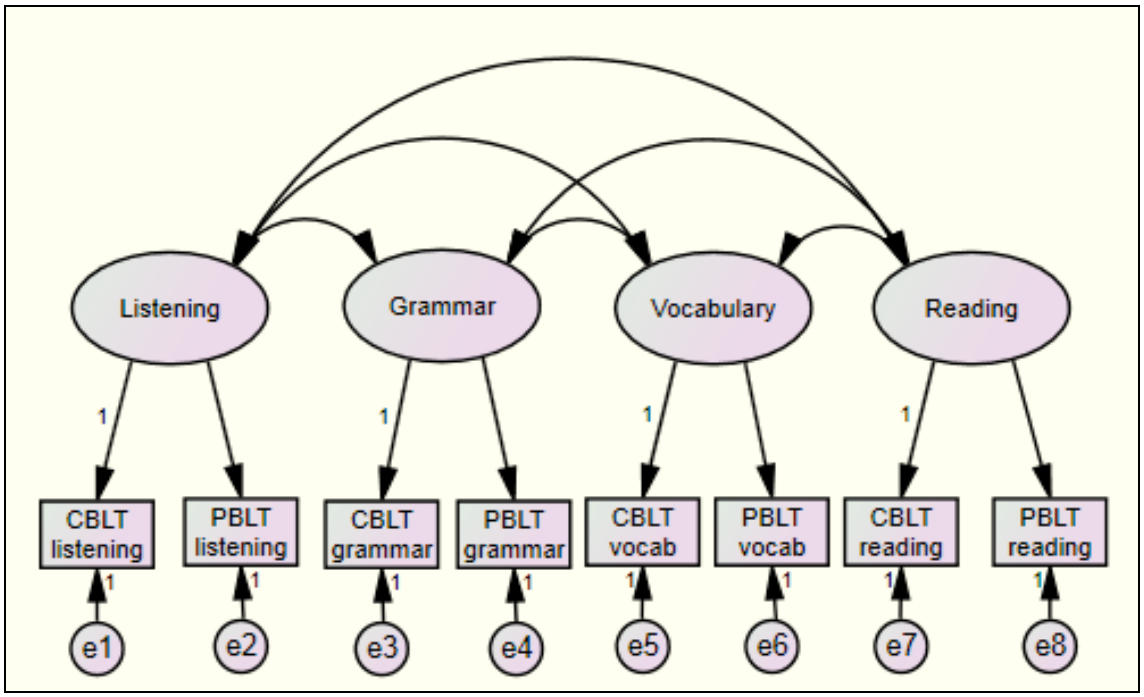


Figure 4. Model specification.

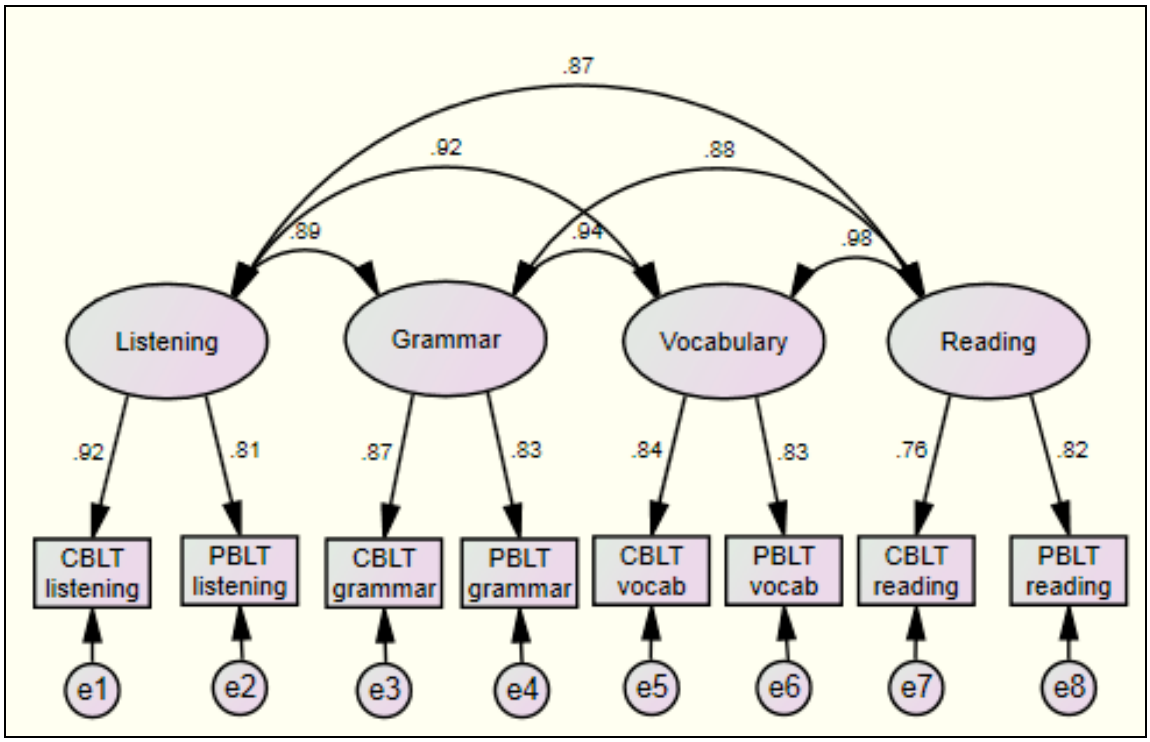


Figure 5. Reproduced model. All parameter estimates are standardized.

All the factor loadings and correlations were identical across the original and reproduced models, as shown in Figures 1 and 5. Fit indices were also successfully reproduced, as seen in Table 1.

Table 1

Fit Indices Across the Original and Reproduced Models

Model	Chi-square	*df*	TLI
Original	33.91	14	0.98
Reproduced	31.81*	14	0.98

Note. TLI = Tucker-Lewis index. **p* = .004. Other fit indices were not reported in the primary study, but were found to be satisfactory in our reproduced model: CFI (Comparative Fit Index) = .988, RMSEA (Root Mean Square Error of Approximation; 90% Confidence Interval) = 0.070 (0.038, 0.103), $p_{\text{close-fit H0}}$ = .137, and SRMR (Standardized Root Mean Square Residual) = .020.

Thus, our illustration shows that we can reproduce SEM results even without the raw data, given access to (1) variances/covariances or (2) correlations and *SD*s (+ means). Our successful reproduction indicates that the model was correctly analyzed and reported in the primary study.

In addition to reproduced models, it is also possible and of great interest to examine alternative models that were not tested in the primary study. Recall that Choi et al. (2003) examined the comparability of a paper-based and a computer-based language test by hypothesizing the four abilities. They examined to what extent the four abilities were equally measured across the two test modes. In this design, it is also possible and would be more sensible to incorporate these two test modes as method factors into an SEM model. The result is shown in Figure 6.

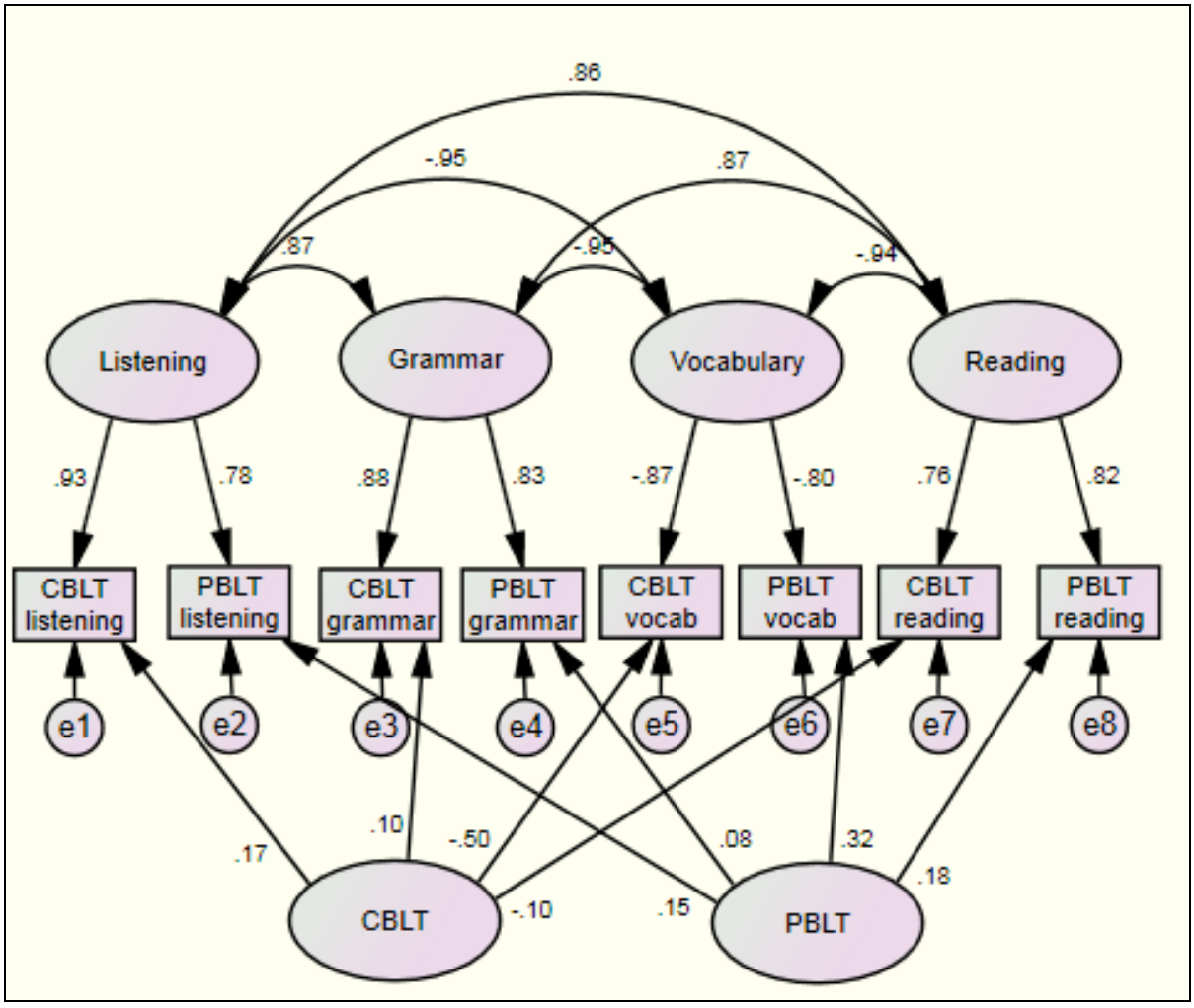


Figure 6. An additional method model. The measurement error variance of the CBLT vocabulary test (e5) was found to be negative but statistically nonsignificant (estimate = –.514, standard error = 6.643, critical ratio = –.077, and p = .938) and thus fixed to zero. If the measurement error variance is negative and statistically significant, the model is unlikely to be adopted (for details, see Chen, Bollen, Paxton, Curran, & Kirby, 2001).

Fit indices in the method model in Figure 6 were satisfactory, χ^2 = 11.394, df = 7, p = .122, CFI = .997, TLI = 0.988, RMSEA (90% CI) = 0.049 (0.000, 0.099), $p_{\text{close-fit H0}}$ = .446, SRMR = .013. Since the models in Figures 1 and 6 are nested, they were compared using a chi-square difference test. The chi-square difference exceeded the threshold value of 14.067 with df = 7, favoring our method model, $\chi^2_{\text{difference}}$ = 33.91 – 11.394 = 22.516, $df_{\text{difference}}$ = 14 – 7 = 7. This suggests that our reproduced method model is statistically better in terms of fit indices. Nevertheless, some of the parameter estimates turned out to be negative in the method model: The correlation between

grammar and vocabulary was –.95, the correlation between vocabulary and reading was –.94, and the factor loadings from the vocabulary factor were –.87 and –.80. They were substantively nonsensical and indicated the inappropriateness of the model. This problem occurred because the estimation of both trait and method factors is often problematic (e.g., Kenny & Kathy, 1992) and/or because the method effect was not strong or suitable to be included in the current model. Thus, Choi et al.'s (2003) model, where the method factors were not incorporated, would be now more strongly supported, given the results that its competing model was not adopted.

Although Choi et al.'s (2003) model was successfully reproduced, reproduction is not always successful. Ellis and Loewen (2007) tested a model of explicit and implicit knowledge, as presented in Figure 7.

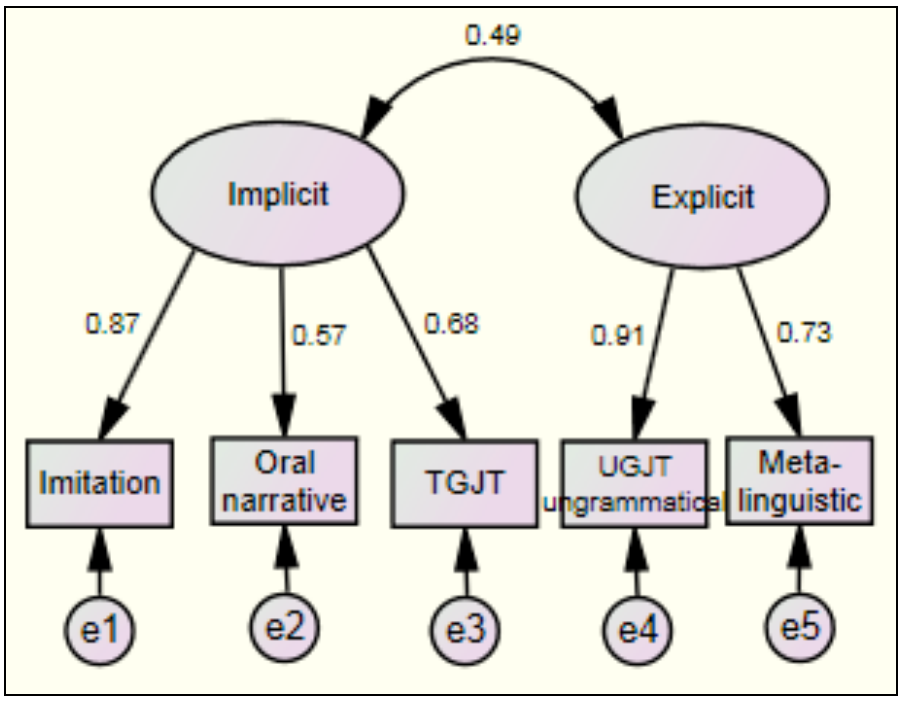
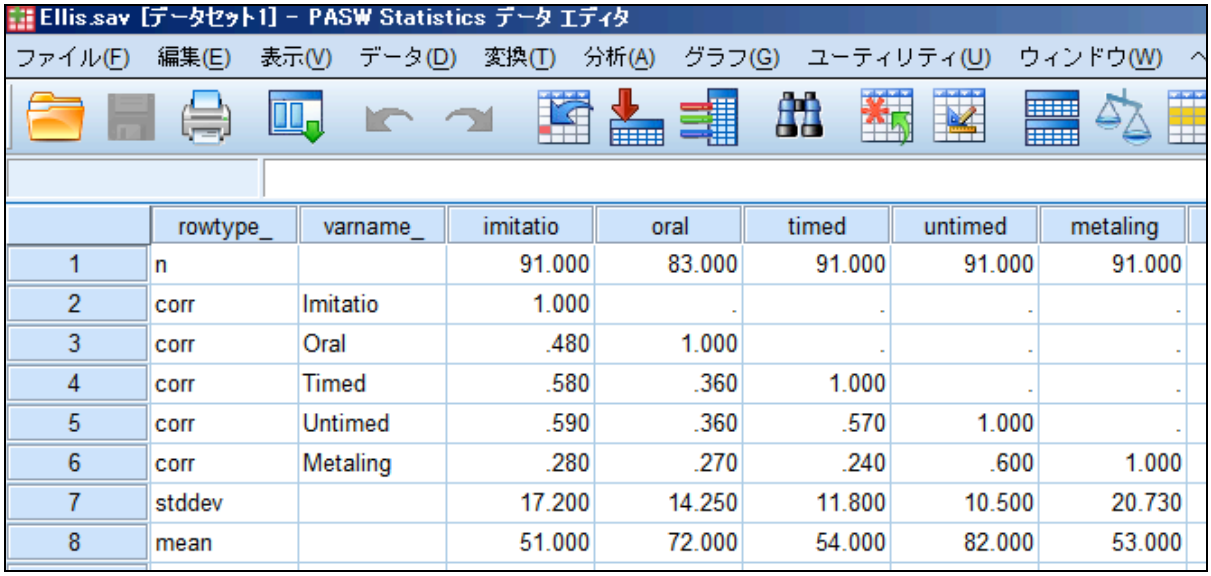


Figure 7. Implicit/explicit model. Adopted from Ellis and Loewen (2007, p. 123). TGJT and UGJT refer to timed and untimed grammar judgment test, respectively. All parameter estimates are standardized.

For our reproduction, the input data were a correlation matrix and *SD*s (see Tables 6 and 7 in Ellis, 2005), as shown in Figure 8. The paths from the implicit factor to the imitation variable, from the explicit factor to the UGJT ungrammatical variable, and from all the measurement errors to the observed variables were fixed to one for identification.



	rowtype_	varname_	imitatio	oral	timed	untimed	metaling
1	n		91.000	83.000	91.000	91.000	91.000
2	corr	Imitatio	1.000	.	.	.	.
3	corr	Oral	.480	1.000	.	.	.
4	corr	Timed	.580	.360	1.000	.	.
5	corr	Untimed	.590	.360	.570	1.000	.
6	corr	Metaling	.280	.270	.240	.600	1.000
7	stddev		17.200	14.250	11.800	10.500	20.730
8	mean		51.000	72.000	54.000	82.000	53.000

Figure 8. SPSS input with the correlation matrix, means, and *SD*s.

The reproduced model is presented in Figure 9.

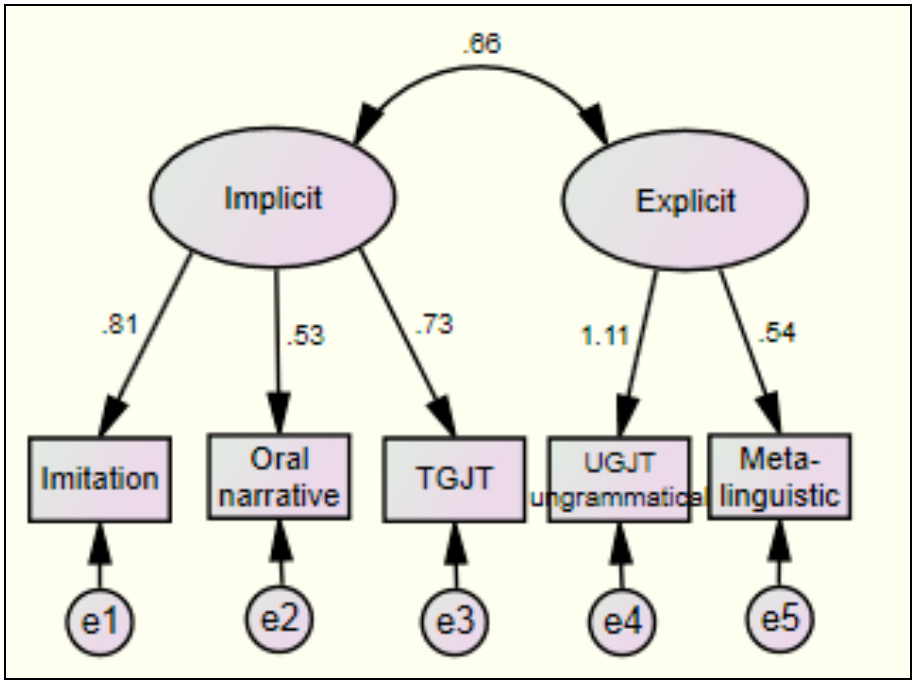


Figure 9. Reproduced model (negative variance e4 unfixed). All parameter estimates are standardized.

The model in Figure 9 was obviously problematic. The factor loading from the explicit knowledge to the untimed grammatical judgment test (UGJT) exceeded 1.00. Furthermore, although it is not presented in Figure 9, the measurement error variance of the UGJT (e4) was found to be negative. Since it was not statistically significant (estimate = –26.183, standard error = 28.267, critical ratio = –.926, and p = .354), the measurement error variance was fixed to zero, on the basis of Chen et al. (2001). The reestimated model is presented in Figure 10.

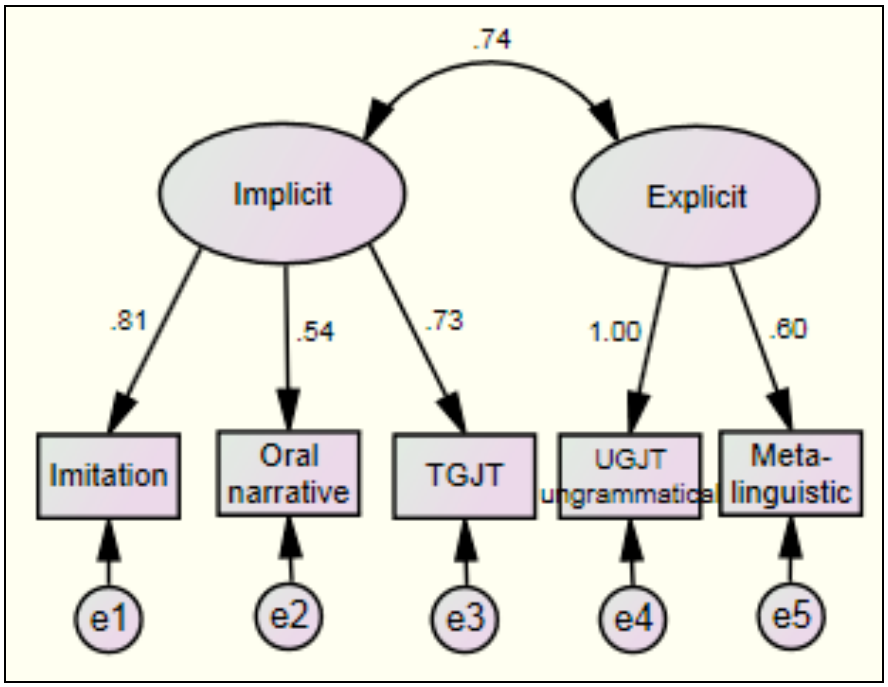


Figure 10. Reproduced model (negative variance e4 fixed). All parameter estimates are standardized.

A comparison of the original and reproduced models in Figures 7 and 10 shows that the reproduction was not successful. For example, parameter estimates were not well reproduced, particularly for the explicit factor: The two factor loadings to the UGJT and metalinguistic test were .91 and .73 in the original study but 1.00 and .60 in our reproduced model. The correlation between the implicit and explicit factors was originally .49 but was reproduced to be .74. Although no guidelines exist for parameter values to be similar before reproduction is considered successful, the divergences observed here were noticeable and indicated that the model reproduction was not as successful as that of Choi et al. (2003). Additionally, the negative error variance was not reported in Ellis and Loewen (2007). It is not clear whether the negative error

variance was obtained in the primary study and, if it was, how it was treated. The failure in reproduction was also confirmed by the fact that three of the first author's graduate students worked independently on a reproduction of the model and all three failed to do so successfully. Thus, although the fit indices in Table 2 suggest the Ellis and Loewen's model was successfully replicated and satisfactory, the divergences in parameter estimates warn of the veracity of—and danger in building a new study based on—their model.

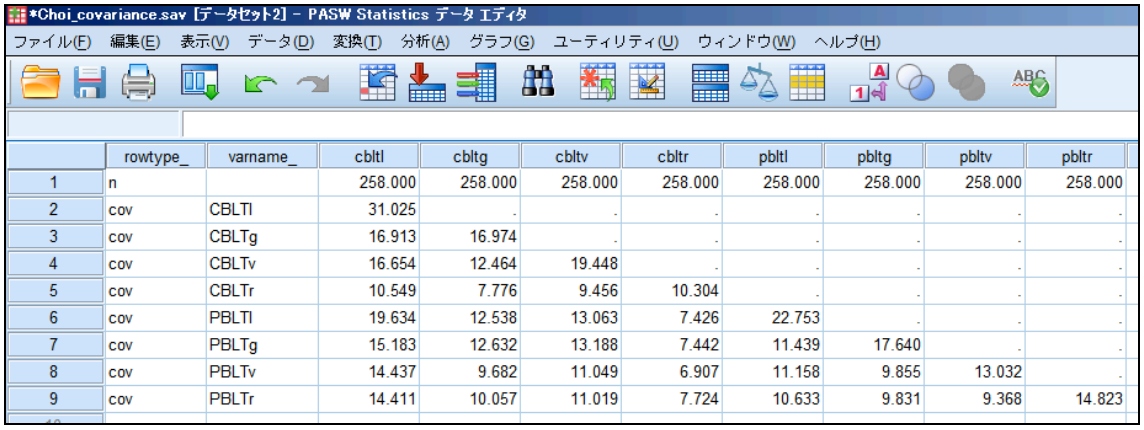
Table 2

Fit Indices Across the Original and Reproduced Models

Model	Chi-square	*df*	NFI	RMSEA	
Original	1.191	4	.991	0.000	
Reproduced (negative variance unfixed)	4.599[a]	4	.971	0.041	(0.000, 0.169)
Reproduced (negative variance fixed)	5.989[b]	5	.962	0.047	(0.000, 0.160)

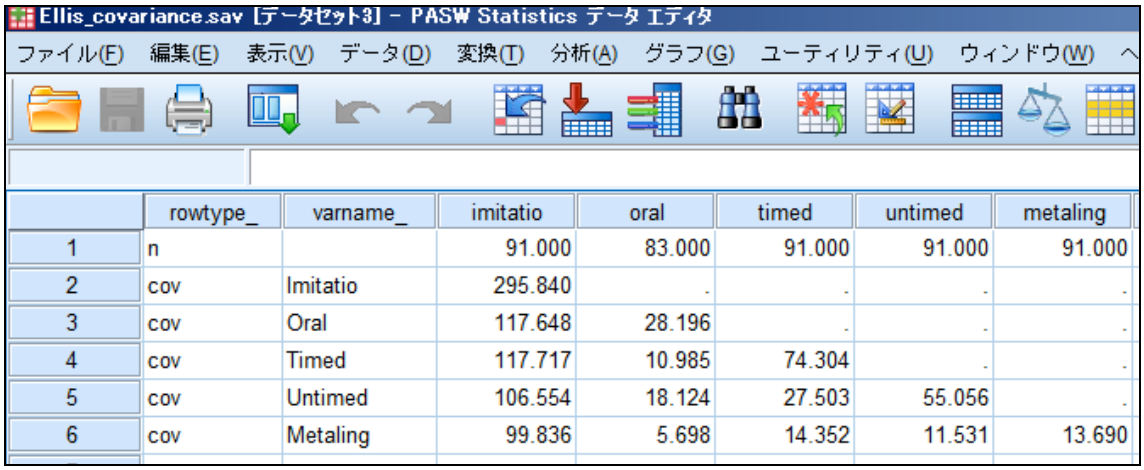
Note. NFI = Normed Fit Index. [a]p = .331. [b]p = .307. Other fit indices were not reported in the primary study but were found to be satisfactory in our reproduced model: CFI = .996, $p_{\text{close-fit H0}}$ = .444, and SRMR = .029 for the reproduced model without the negative variance fixed; CFI = .993, $p_{\text{close-fit H0}}$ = .433, and SRMR = .036 for the reproduced model with the negative variance fixed.

The demonstration so far has been based on correlation matrices with *SD*s. If you have variance/covariance matrices, the input file should look similar to Figures 11 and 12. Since (a) variance/covariance matrices and (b) correlation matrices with *SD*s are mathematically equal, the results should be exactly the same.



	rowtype_	varname_	cbltl	cbltg	cbltv	cbltr	pbltl	pbltg	pbltv	pbltr
1	n		258.000	258.000	258.000	258.000	258.000	258.000	258.000	258.000
2	cov	CBLTl	31.025	.	.	.	.	.	.	.
3	cov	CBLTg	16.913	16.974	.	.	.	.	.	.
4	cov	CBLTv	16.654	12.464	19.448	.	.	.	.	.
5	cov	CBLTr	10.549	7.776	9.456	10.304	.	.	.	.
6	cov	PBLTl	19.634	12.538	13.063	7.426	22.753	.	.	.
7	cov	PBLTg	15.183	12.632	13.188	7.442	11.439	17.640	.	.
8	cov	PBLTv	14.437	9.682	11.049	6.907	11.158	9.855	13.032	.
9	cov	PBLTr	14.411	10.057	11.019	7.724	10.633	9.831	9.368	14.823

Figure 11. SPSS input with the variance/covariance matrix of Choi et al. (2003). Diagonal values show variances (calculated by a square of *SD*). Other values indicate covariances (calculated by *r* multiplied by *SD* of one variable and *SD* of the other variable).



	rowtype_	varname_	imitatio	oral	timed	untimed	metaling
1	n		91.000	83.000	91.000	91.000	91.000
2	cov	Imitatio	295.840	.	.	.	.
3	cov	Oral	117.648	28.196	.	.	.
4	cov	Timed	117.717	10.985	74.304	.	.
5	cov	Untimed	106.554	18.124	27.503	55.056	.
6	cov	Metaling	99.836	5.698	14.352	11.531	13.690

Figure 12. SPSS input with the variance/covariance matrix of Ellis and Loewen (2007).

4. Frequently Asked Questions

After understanding the procedure for replicating the original models, readers may have the following five questions, which we attempt to answer accordingly.

(1) To what extent can the original models be successfully replicated?

(2) Why do we obtain different results between the original model and reproduced model?

(3) When the article does not report complete information necessary for reproduction, can we still reproduce the model?

(4) When we contact the authors, how likely is it for us to obtain the information not reported in the original article?
(5) What should we do to improve data sharing?

4.1 To What Extent Can the Original Models Be Successfully Replicated?

As seen above in the example of Ellis and Loewen (2007), differences are observed in fit indices and parameters between the original and the replicated models. How often do we encounter such divergence? In'nami and Koizumi (2010) aimed to answer this question by collecting previous studies using SEM in the field of second language testing and learning, reproducing those models, and synthesizing the results.

We searched for studies published in 20 international journals using SEM in October 2008 and identified 50 such articles. By retrieving the information necessary for reproduction from the articles and from the authors, we collected such information from 23 of the 50 articles (124 models of the 360 models). We examined if we could replicate the same models tested by the primary researchers; we did not test competing models that the original authors did not test. Table 3 shows the breakdown of the models.

Table 3
Input Information for Reproduction

Variances, covariances, correlations, means, and *SD*s for 9 models
Variances and covariances for 38 models[a]
Correlations, means, and *SD*s for 47 models[b]
Correlations for 30 models

Note. Not reported in In'nami and Koizumi (2010) owing to space limitations. A list of the models we gathered is available upon request. [a]including such information for 16 models obtained from two authors. [b]including such information for one model obtained from one author and for 10 models from another author.

We found that in the preliminary analysis, 89% of the models were successfully replicated. In other words, we mostly obtained the same results as the original articles. When we examined fit indices, 87% to 100% of the analyzed models showed very similar results, depending on the fit index used. In terms of parameter estimates, we generated very similar results for 94% of the models. In conclusion, the results overall suggest that most of the models reported in previous studies were successfully

replicated, indicating that the findings that were obtained through SEM in second language testing and learning are in most cases credible, allowing them to form the basis for future research.

We also compared the replication rate with Hulland et al. (1996), who conducted a similar replication study in marketing. They considered a model to be successfully replicated when differences in (probably standardized) parameter estimates between the original and the replicated results were within 0.10. Table 4 shows the similar replication rates between their results (67%) and ours (57%). This indicates a similar, moderate degree of model replicability across domains.

Therefore, the answer to the question is as follows. The original models can be successfully replicated to a moderate or large degree (57% to 100%), depending on the criteria or aspects focused. Since the successful replication rate was not always perfect, we suggest reanalyzing the model whenever possible and checking to see if similar results can be derived before interpreting the model or starting new research based on the model.

Table 4

Comparison of Replication Rate

	Total number of models	Models lacking information for replication	Models having analytic problems[a]	Models successfully replicated[b]	Models un-successfully replicated[b]
Current study	360	236 + 70[c]	19	20 (57%)	15 (43%)[d]
Hulland et al. (1996)	343	231	--	75 (67%)	37 (33%)

Note. Not reported in In'nami and Koizumi (2010) owing to space limitations. -- = not reported. [a]Examples are problems with model unidentification or with not positive definite matrices. [b]The criterion for successful replication is based on Hulland et al. (1996). [c]These 70 models did not report parameter estimates although we were able to retrieve the correlation or the covariance matrices. [d]In addition to these 15 models, there were other 5 models that had the parameter estimates reported in the original studies but had analytic problems during the replication; these 5 models were classified as "Models having analytic problems."

4.2 Why Do We Obtain Different Results Between the Original Model and Reproduced Model?

Although we occasionally see differences between the original and the replicated models, identifying the exact cause of this is usually difficult. If the same differences are observed even after we double check the input and output of the reproduced models, there are three other possible reasons. A first possible cause of unsuccessful reproduction is that the original models were in some way incorrectly reported. A second reason is that differences exist due to rounding, employing different software programs, and other details that are not reported in the primary studies (e.g., estimation methods). For example, software programs calculate chi-squares slightly differently (e.g., Schumacker & Lomax, 2004).

Third, we may not obtain exactly the same results as those reported in the published articles if the original data did not satisfy univariate and/or multivariate normality—one of the assumptions required for SEM—and when primary researchers used special adjustment techniques for such data (e.g., using Satorra-Bentler-scaled chi-square statistics or bootstrapping estimates of parameters for nonnormal data). The same applies when the primary data had missing values and when researchers used special procedures for it. In these cases, we cannot make the same adjustment as in the primary studies without raw data. This prevents us from reproducing the model, although close reproduction may be possible if nonnormality and/or missingness are less severe.

Although it is often difficult to identify the precise reasons of divergence of the models, we argue that reproduction of the original models is useful in providing researchers with the opportunity to critically appraise the adequacy of the models.

4.3 When the Article Does Not Report Complete Information Necessary for Reproduction, Can We Still Reproduce the Model?

If we have only *SD*s and do not have correlation matrices, reproduction is not possible. If we have correlation matrices and do not have *SD*s, reproduction is possible by inputting 1.000 for *SD*s and following the same procedure as we do when we have both correlation matrices and *SD*s. However, (a) correlation matrices with *SD*s or (b) variance/covariance matrices are more preferable to (c) correlation matrices without *SD*s, because reanalysis using (c) generates imprecise standard errors of parameters, which results in incorrect *p*-value for the parameters (Cudeck, 1989). Thus, if you are

interested in reproducing statistical significance of parameters, this method is not recommended.

4.4 When We Contact the Authors, How Likely Is It for Us to Obtain the Information Not Reported in the Original Article?

When the information required is missing, we can contact the authors. According to In'nami and Koizumi (2010), however, they do not always respond to such inquiries. Of the 50 SEM articles we retrieved (see section 4.1), only 19 articles (38%) had sufficient information for reanalysis. Since 31 articles failed to report such information, we sent a query to the 33 authors of these 31 articles. The e-mail message we sent included the purpose of our study, which was, of course, to reproduce the original models to check if the same results are observed. If we obtained a reply, we further asked why the information was not included in the article.

As seen in Table 5, we obtained responses from 18 authors but successfully obtained information necessary for reanalysis from only four authors (12%). Interestingly, three of the four authors published in *Language Testing*, suggesting that second language testers may be more willing to share data. Therefore, the answer to the question of how likely it is for us to obtain the information not reported in the original article is that we can expect the odds of the successful retrieval of data to only be approximately one out of 10.

Table 5

Response Frequency From Article Authors

Reply	18
	Information obtained = 4
	Variance/covariance matrix = 2
	Correlation matrix with means and *SD*s = 1
	Raw data = 1
	Information not obtained = 14
	Data too old to be retrieved = 6
	No further response despite promise = 6
	Discarded data when moving, as the data were decades old and occupied much space in the office = 1
	Tried to send the raw data which were too large to cause a transmission problem and did not reach us = 1
	Reasons for not reporting the information in the primary articles = 8
	They were not asked to do so by the editor(s) or reviewers = 4
	They used raw data to deal with missing data problems and thought it was not helpful to report the matrices = 4
	They could not report owing to space limitations = 2
	They accidentally failed to include such information = 1
	It was not reported in previous studies either = 1
No reply	13

Note. Not reported in In'nami and Koizumi owing to space limitations (2010).

Therefore, the percentage of obtained data required for reproduction from authors (12%) was found to be very low. Our success rate for obtaining the data from the authors was less than, but similar to, previous studies in which the success rate for retrieving psychological raw data from the authors was 24% (Wolins, 1962), 26% (Wickerts, Borsboom, Kats, & Molenaar, 2006), and 38% (Craig & Reese, 1973).

Although we do not know exact reasons why authors did not respond to our request to share the information, one reason might be that data sharing and replication are rare among second language researchers. Many second language researchers may have little experience of sharing data or information not written in the article, may be less concerned about sharing, or may not be very aware of the value of replication. In

fact, among the 50 articles containing a total of 360 models we reviewed, none reported attempts to replicate models tested by primary researchers.

Further, Wickerts et al. (2006) stated that authors are reluctant to share data probably because of the time and effort needed to prepare a data file and because of the lack of benefits they would receive in return for their time and effort. Other reasons, according to King (1995), would be that when data are derived from funded projects, authors may not have the right to send the data for reanalysis to those who request it, and that a study that cannot be replicated would lead to criticism of the original finding. While Wickerts et al. (2006) and King (1995) described the cases when raw data were requested for reanalysis, the current study asked for (a) the variance/covariance matrix or (b) the correlation matrix along with *SD*s. We initially expected that providing such matrices would be more manageable and would not require the original authors to spend as much time and effort to send than sending raw data would. However, we learned the hard way that in this field, even retrieving matrices is difficult. It seems that although the American Psychological Association (2010) guidelines are followed by many journals, the following principle on sharing data for reanalysis is not well known or widely practiced.

> Once an article is published, researchers must make their data available to permit other qualified professionals to confirm the analyses and results (APA Ethics Code Standard 8.14a, Sharing Research Data for Verification). Authors are expected to retain raw data for a minimum of five years after publication of the research. Other information related to the research (e.g., instructions, treatment manuals, software, details of procedures, code for mathematical models reported in journal articles) should be kept for the same period; such information is necessary if others are to attempt replication and should be provided to qualified researchers on request (APA Ethics Code Standard 6.01, Documentation of Professional and Scientific Work and Maintenance of Records). (American Psychological Association, 2010, p. 12)

In the end, the information necessary for replication was obtained from 23 (with 124 models) of the 50 articles (with 360 models). Therefore, the overall success rate of retrieving the necessary information for reproduction was 34% (124/360). This low rate was similar to that in marketing (33%; 112 of 342 models, in Hulland et al., 1996), suggesting that only one-third of the researchers in second language testing and

learning and in marketing make information for replication available.

4.5 What Should We Do to Improve Data Sharing?

So far, we have seen predicaments in which most researchers are hesitant to share their data with those interested in reproducing their models and where the information necessary for reproduction is not often available to external researchers. To alter this situation, we propose two feasible ways. First, journal editors and reviewers should check that (a) the variance/covariance matrix and/or (b) the correlation matrix along with means and *SD*s are reported for any SEM analysis at the reviewing stage. If space limitations prevent editors from placing such information in the journals, their websites can hold the information as additional material, such as is found in *Applied Linguistics* and *Language Learning*. Moreover, authors can post such information on their website. Since (a) is reproducible from (b), and means are not always necessary for replication but aid data interpretation, we recommend reporting (b). If analysis is accompanied by syntax, model replication can be further enhanced. We also recommend archiving raw data for replication (see American Psychological Association, 2010; see King, 1995 for issues of data confidentiality). Such practice can be especially viable if it is encouraged by journal editors and reviewers, because they play an important role in helping authors determine whether such information should be reported (see Table 5).

When data are nonnormally distributed and/or included missing data, and successful replication is difficult without the raw data, this can be stated in articles as a reason for not providing the variance/covariance or correlation (with the *SD*s) matrix. It would still be useful to report matrices based on the listwise deletion of missing data, even when the authors use pairwise deletion of missing data or full information maximum likelihood estimation, whereby a model with missing data is analyzed as it is and reanalysis requires raw data.

Second, we should encourage more replication in general, and particularly the reproduction of structural equation models in second language testing and learning. Researchers should conduct reanalysis of structural equation models in their research, critically examine them, and not take them at face value. Furthermore, model reanalysis also plays a role in the introduction part of a new study: Researchers can report the results of replicating existing models, critically evaluate the previous findings, and strengthen the need to conduct the new (or further replicated) study. They can even test any competing models that were not initially performed in primary studies. Such reanalysis has substantial educational value because it can send strong

messages to researchers, especially young ones: the importance of replication and reanalysis in general, the need to confirm the credibility of models reported in previous studies by replication, the risks of building on previous work without close scrutiny, the necessity of correctly reporting important and necessary information for later reanalysis, and the importance of responding to inquiries about their own work.

We hope that these routine practices for data accessibility and model replication will increase the openness of scientific research and help us accumulate more knowledge within the field.

Acknowledgements

This research was partially supported by a Grant-in-Aid for Scientific Research (KAKENHI) from the Ministry of Education, Culture, Sports, Science and Technology in Japan (No. 23720283).

References

Abbuhl, R. (2012). Why, when, and how to replicate research. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 296–312). West Sussex, U.K.: Wiley-Blackwell.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Arbuckle, J. L. (1994–2011). Amos [Computer software]. Chicago, IL: Smallwaters.

Arbuckle, J. L. (2009). Amos (Version 18.0.0) [Computer software]. Chicago, IL: Smallwaters.

Bentler, P. M. (1994–2011). EQS for Windows [Computer software]. Encino, CA: Multivariate Software.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.

Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Erlbaum.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378–399.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper

solution in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research*, *29*, 468–508.

Choi, I.-C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing*, *20*, 295–320.

Craig, J. R., & Reese, S. C. (1973). Retention of raw data: A problem revisited. *American Psychologist*, *28*, 723.

Dandonoli, P., & Henning, G. (1992). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals*, *23*, 11–22.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172.

Ellis, R., & Loewen, S. (2007). Confirming the operational definitions of explicit and implicit knowledge in Ellis (2005): Responding to Isemonger. *Studies in Second Language Acquisition*, *29*, 119–126.

Fienberg, S. E., Martin, M. E., & Straf, M. L. (1985). *Sharing research data*. Washington, DC: National Academy Press.

Finifter, B. M. (1975). Replication and extension of social research through secondary analysis. *Social Science Information*, *14*, 119–153.

Fitzpatrick, T. (2009). Integrating replication work into a PhD programme. In G. Porte (Chair), *Encouraging replication research in the field of AL and SLA*. Colloquium conducted at the American Association for Applied Linguistics Annual Conference, Denver CO.

Freese, J. (2007). Replication standards for quantitative social science: Why not sociology? *Sociological Methods & Research*, *36*, 153–172.

Fulcher, G. (1996). Invalidating validity claims for the ACTFL Oral Rating Scale. *System*, *24*, 163–172.

Gass, S. M., & Mackey, A. (2007). *Data elicitation for second and foreign language research*. Mahwah, NJ: Erlbaum.

Gorsuch, G. J. (2000). EFL educational policies and educational cultures: Influences on teachers' approval of communicative activities. *TESOL Quarterly*, *34*, 675–710.

Hedges, L. V. (1987). How hard is hard science, How soft is soft science? The empirical cumulativeness of research. *American Psychologist*, *42*, 443–455.

Hulland, J., Chow, Y.-H., & Lam, S. (1996). Use of causal models in marketing research: A review. *International Journal of Research in Marketing*, *13*,

181–197.

In'nami, Y. (2011). *An introduction to structural equation modeling for vocabulary research*. Invited workshop held at the 8th JACET (Japan Association of College English Teachers) Vocabulary Acquisition Research Group Conference, Tokyo, Japan. Retrieved from https://e-learning.ac/jlta.ac/file.php/1/111210SEM_workshop_Innami.pdf

In'nami, Y., & Koizumi, Y. (2010). Can structural equation models in second language testing and learning research be successfully replicated? *International Journal of Testing, 10,* 262–273. doi: 10.1080/15305058.2010.482219

In'nami, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly,* 8, 250–276. doi:10.1080/15434303.2011.565844

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, *112*, 165–172.

King, G. (1995). Replication, replication. *PS: Political Science and Politics*, *28*, 443–451.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

La Sorte, M. A. (1972). Replication as a verification technique in survey research: A paradigm. *Sociological Quarterly*, *13*, 218–227.

Language Teaching Review Panel. (2008). Replication studies in language learning and teaching: Questions and answers. *Language Teaching*, *41*, 1–14.

Lee, S.-y. (2005). Facilitating and inhibiting factors in English as a Foreign language writing performance: A model testing with structural equation modeling. *Language Learning*, *55*, 335–374.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151–159.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Erlbaum.

Muthén, L. K., & Muthén, B. O. (1998–2011). Mplus [Computer software]. Los Angeles, CA: Muthén & Muthén.

Ortega, L. (2009, March). A sociology of replication and replicability in applied linguistics. In G. Porte (Chair), *Encouraging replication research in the field of*

AL and SLA. Colloquium conducted at the American Association for Applied Linguistics Annual Conference, Denver CO. Retrieved May 31, 2009, from http://www2.hawaii.edu/~lortega/Ortega2009replication.ppt

Polio, C., & Gass, S. (1997). Replication and reporting: A commentary. *Studies in Second Language Acquisition*, *19*, 499–508.

Porte, G. (2009, March). *Encouraging replication research in the field of AL and SLA*. Colloquium conducted at the American Association for Applied Linguistics Annual Conference, Denver CO.

Porte, G. K. (2010). *Appraising research in second language learning: A practical approach to critical analysis of quantitative research* (2nd ed.). Amsterdam, the Netherlands: John Benjamins.

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modelilng* (2nd ed.). Mahwah, NJ: Erlbaum.

Santos, T. (1989). Replication in applied linguistics research. *TESOL Quarterly*, *23*, 699–702.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling* (2nd ed.). Mahwah, NJ: Erlbaum.

Titscher, S., Meyer, M., Wodak, R., & Vetter, E. (2000). *Methods of text and discourse analysis*. London: Sage.

Valdman, A. (1993). Replication study. *Studies in Second Language Acquisition*, *15*, 505.

Valdman, A. (1997). A word from the editor. *Studies in Second Language Acquisition*, *20*, 67.

Wickerts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728.

Wolins, L. (1962). Responsibility for raw data. *American Psychologist*, *17*, 657–658.