Can Structural Equation Models in Second Language Testing and Learning Research be Successfully Replicated?

Yo In’nami

Toyohashi University of Technology, Japan

Rie Koizumi

Tokiwa University, Japan

Resubmission date: March 28, 2010

Author Posting. (c) Taylor & Francis Group, LLC, 2010.

This is the author's version of the work. It is posted here by permission of 'Copyright Holder' for personal use, not for redistribution. The definitive version was published in International Journal of Testing, Volume 10 Issue 3, July 2010.

doi:10.1080/15305058.2010.482219

Abstract

Because structural equation models are widely used in testing and assessment, investigation into the accuracy of such models may help raise awareness of the value of reanalysis or replication. We focused on second language testing and learning studies, and examined: (a) To what extent is information necessary for replication provided by authors? (b) To what extent can the original models be successfully replicated? Regarding (a), we e-mailed authors of 31 articles that did not contain information needed to replicate the study and asked them for the missing information. We obtained data from only four authors. Regarding (b), we succeeded in replicating 89% of the models in the preliminary analysis, 87-100% of fit indices, and 94% of parameter estimates. The results suggest that for the most part, SEM research reported in second language testing and learning research is accurate.

Keywords: second language testing, second language learning, structural equation modeling, replication
Can structural equation models in second language testing and learning research be replicated?

**Background**

**Replication**

One of the most important issues widely discussed across the academic domain is the replication of findings. There are two primary reasons behind the significance of this topic. First, replication helps empirical and theoretical findings build on one another so that the results of current research can expand upon earlier work (e.g., Hedges, 1987; Kline, 2004). Second, replication encourages researchers to appraise studies more critically, since conducting good replication studies presupposes the proper understanding of the original studies (e.g., King, 1995).

One type of replication that has not been discussed extensively is model replication using structural equation modeling (SEM). Although such replication is common in books on SEM (e.g., Bollen, 1989; Raykov & Marcoulides, 2006), this is not the case in the field of testing and assessment. For example, Furlan, Cassady, and Pérez (2009) reported that the factor structure of a Spanish version of the Cognitive Test Anxiety Scale was unidimensional and consistent with that of the original English version. However, they did not conduct a replication study on the English version to ensure that its factor structure was indeed unidimensional. The situation is especially problematic in testing and assessment, where
SEM is widely used to assess the construct validity of instruments, but replication studies are rarely conducted.

Further, while replication of structural equation models can be conducted for each previous study, of greater interest would be a synthesis of a replication of structural equation models. In marketing, Hulland, Chow, and Lam (1996) collected 343 models using SEM in 186 articles and found that only 112 out of the 343 models (33%) reported sufficient information for reanalysis. They successfully replicated 75 out of the 112 models (67%). Hulland et al. partially attributed this to errors in reporting in the original studies and suggested risks in building future research on previous work. It should be noted that these authors did not solicit the necessary information for reanalysis from the authors of the original studies, because differences in response rates could bias the results. However, their approach, as Hulland et al. admitted, could have underestimated potential replicability, because interested researchers can contact the authors of the original studies and likely obtain necessary information, especially if the studies were conducted within five years after the date of publication, the minimum required period for retention of the original data (American Psychological Association, 2001).

Data Availability for Replication

Since replication of structural equation models requires the variance/covariance or correlation matrices, we need to obtain such information by reading articles or contacting the
authors. One question is to what extent the data are available from the authors. The American Psychological Association (2001) states that “psychologists do not withhold the data . . . from . . . professionals who seeks to verify the substantive claims through reanalysis” (p. 396). However, Wickerts, Borsboom, Kats, and Molenaar (2006) could obtain data from only 64 of the 249 studies (26%) that appeared in the last two 2004 issues of four major APA journals after they contacted the article authors. This success rate was similar to that reported by Wolins (1962, 24%) and Craig and Reese (1973, 38%). Wickerts et al. posited that authors’ reluctance to share data is probably due to the meager incentives authors receive in return for time and effort needed to prepare a manageable data file.

Current Study

We investigated two research questions: (a) To what extent is information necessary for replication provided by contacting authors? (b) To what extent can the original models be successfully replicated? Both questions were examined with special focus on second language testing and learning studies.

Method

Article Collection

Studies using SEM in second language testing and learning were collected in October 2008, in two ways. First, studies in 20 representative journals (e.g., Applied Linguistics,
Analyses

Research Question 1

We extracted information necessary for replication (i.e., a variance/covariance matrix or a correlation matrix with means and SDs) from articles using SEM. When such information was missing, we e-mailed all the (co-)authors and requested the matrix to investigate accessibility of information for replication. In the e-mail, we explained the purpose of our research (i.e., reanalysis of the original models to verify if the same results are obtained), and if we received the data we asked why the matrix was not included in the article. When we did not receive a reply, we did not send a reminder.

Research Question 2

Model replication was based on (a) variance/covariance matrices, (b) correlation matrices with SDs, or (c) correlation matrices without SDs. (a) and (b) are mathematically

1 Further details in this regard and other parts of the current paper are available upon request.
equal and are more preferable to (c), because reanalysis using (c) produces incorrect standard errors of parameters and hence leads to the incorrect level of statistical significance for the parameters (Cudeck, 1989). However, we also used (c) because we aimed to reproduce fit indices and standardized parameter values, not standard errors of parameters or significance levels of parameters.

Unavailability of the raw data precluded examination of univariate and multivariate normality, use of the Satorra-Bentler scaled chi-square ($\chi^2$) statistic or bootstrapping estimates of parameters in (plausibly) non-normal data, and treatment of missing data.\(^2\) When there were multiple models in one study, each model was replicated: Not only the final models but also all the models tested and reported in the study were included and reanalyzed.

We replicated each model with Amos 4.01 (Arbuckle, 1999), by exactly following the procedures reported in the original articles. When these procedures were not reported in detail, replication was conducted using maximum likelihood estimation methods. One of the factor loadings from each factor was fixed to 1 for scale identification.

For model replication, we focused on (a) preliminary analysis (non-convergence and negative error variances), (b) model fit indices, and (c) parameter estimates (i.e., factor loadings, correlations, and error variances). Regarding (a) preliminary analysis, when

---
\(^2\)The 124 models reanalyzed (see section III.2) did not include models with missing data where full information maximum likelihood, multiple imputation, or some imputation methods (not specified in the original articles) were used to address missing data. We failed to obtain correlation/covariance matrices for these models (i.e., five articles containing 32 models in total). Thus, the results of our replication were not affected by existence of missing data in the original models.
non-convergence to proper solutions occurred within 500 iterations (the default value for Amos 4.01), the parameter values from the last iteration were used as the starting values for a new series of iterations (Bollen, 1989). When a model still did not converge, a start value of 0.1, 0.3, or 0.5 was used for parameters and a start value of 100 for variances (Kline, 2005; Raykov & Marcoulides, 2006). Statistically nonsignificant negative error variances were fixed to zero, and the models were reestimated (Chen, Bollen, Paxton, Curran, & Kirby, 2001).

Once a model converged and had no statistically significant negative error variance(s), and all the parameter estimates were confirmed within the expected range, we calculated the absolute value differences between each original model and its replicated one in (b) model fit indices and (c) parameter estimates, in order to examine how exactly the model was replicated. For example, if one of the fit indices, RMSEA, was 0.030 for the original model and 0.040 for our replicated model, the absolute value difference of 0.010 was coded. Regarding parameter estimates, since initial analyses showed that parameter estimates for each original and replicated model differed little, we focused on the largest different parameter values for each model.

We also counted the number of models whose absolute value differences in model fit indices and parameter estimates between the original and replicated models lay out of the mean +/- 2SD range (thus judged to be outliers, or unsuccessful replication). For example,
when a mean absolute value difference in NFI between the original and replicated models was 0.036 with an SD of 0.052, the number of models whose value differences lay out of −0.068 and 0.140 (0.036+/− 2*0.052) was counted (see Table 1). It should be noted that because we used the absolute differences, having a confidence interval going beyond zero (−0.068) is theoretically possible but practically meaningless. Thus, the actual range within which the model was considered to be successfully replicated was zero or above (e.g., 0.000 to 0.140 for NFI).

Results and Discussion

Research Question 1: Availability of Information for Replication from the Authors

We found 50 articles using SEM in 7 of the 20 journals we reviewed. Nineteen articles out of 50 (38%) contained enough information for reanalysis. We e-mailed the 33 authors of these remaining 31 articles missing information for reanalysis. Eighteen authors responded, but only 4 of them (12% [4/33]) provided information. Among the 31 articles, published within the last five years (2003 to 2008) were 22 of the 33 (71%), including 11 of the 18 articles whose authors responded and 11 of the 13 articles whose authors did not respond.

Asked why the information necessary for reanalysis was not reported in the original articles, the 4 authors who sent us their data said: they were not asked to do so by the editor(s) or reviewers (4 authors), they could not report due to paper length (2 authors), they simply failed to report (1 author), and they did not report because such information was not
reported in previous studies (1 author). Thus, the editor(s) or reviewers may be influential in the authors’ decision whether to include information for reanalysis.

Eventually, information necessary for reanalysis was obtained from the articles and the authors for 23 (with 124 models) out of the 50 articles (with 360 models) in the 7 journals. Those 50 articles are marked with asterisks (*) in the references, among which the 23 articles that provided information for reanalysis are marked with double asterisks (**). Thus, the rate of successfully obtaining the necessary information for reanalysis was 34% (124/360).

**Research Question 2: Replicability of the Original Models**

We reanalyzed the 124 models with sufficient information for reanalysis. During (a) preliminary analysis, 19 of the 124 models faced problems with either model unidentification (13 models) or not positive definite matrices (6 models). These problems were also observed in original analyses in 3 of the unidentified models and 2 of the models with non-positive definite matrices. Thus, the 10 unidentified models and the 4 models with not positive definite matrices did not have such problems in the original study but did in our replication, suggesting that 11% (14/124) of the models reanalyzed failed to be replicated. In other words, 89% of the models were successfully replicated in the preliminary analysis.

For the remaining 105 models without analytical problems, the differences in values of (b) fit indices and (c) parameters were examined to investigate to what extent the original models and replicated models were similar. Table 1 shows the descriptive statistics of
absolute value differences of fit indices and parameters between the original and replicated models. Since skewness and kurtosis were overall large, we based our interpretation on medians, rather than means. The results show that the differences in values of fit indices between the original studies and our reanalysis were, in most cases, very small (see the near-zero values of median and the large positive values of skewness and kurtosis). Further, the number of models whose absolute value differences in model fit indices indicated as outliers was small (0 to 13%). In other words, successful replication rate was high, ranging from 87% to 100%, depending on the fit index used. For example, an absolute value difference of NFI between the original and replicated models was small (median = 0.015), distributed near zero (skewness = 2.214), and had a strongly peaked distribution (kurtosis = 5.670), with 1 of the 24 models judged to be an outlier (4%).

[Insert Table 1 about here]

Regarding parameter estimates, Table 1 shows that parameter estimates differed little between the original and replicated models (median = 0.040), with most estimates distributed around the mean (skewness = 2.284; kurtosis = 5.529). Only 2 out of the 35 parameters (6%) failed to be replicated, suggesting the successful replication rate of 94%. The largest difference in parameter values was 1.780, where a parameter estimate sign flipped between the original (0.89) and replicated (–0.89) models although the remaining parameters for this model were well replicated.
Hulland et al. (1996) replicated models in marketing and regarded a model as successfully replicated when (probably standardized) parameter estimates varied by 0.10 or below between the published and the replicated results. Based on their criterion, 20 out of the 35 models in our reanalysis (57% [20/35]) had the largest difference of 0.10 or below in parameter estimates between the original and replicated models. Thus, successful replication rates between the Hulland et al. in marketing (67%) and the current study in second language testing and learning studies (57%) were similar, suggesting some degree of generalizability of model replicability across domains.

Despite the low percentage of outlier models, some of the outliers were noticeable and problematic. Some divergences are expected due to rounding, software differences (e.g., Schumacker & Lomax, 2004), and other details that were not reported in the original studies (e.g., estimation methods). However, the chi-square differences in two outlier models (1601.104 and 6932.187) were too large to be explained by these reasons. Further, degrees of freedom (df) should be the same under the same model specification. Nevertheless, one model had the maximum difference of 66 in degrees of freedom between the original and replicated models.

Although identifying the cause of discrepancies in fit indices and parameters between the original and the replicated models is difficult, we explored possible reasons. For the 25 unsuccessfully replicated models as found by totaling the number of models in the rightmost
column in Table 1, these were actually a total of 17 models since the same models failed to be replicated in more than one fit indices and/or parameters and were counted accordingly. Of the 17, only two models were analyzed after univariate and multivariate normality were checked, and only one model was analyzed after missing data was handled with listwise deletion methods; issues on data normality and missing data were not reported for the remaining models. Accordingly, non-normality and missing data might have caused unsuccessful replication.

Other possible causes of unsuccessful replication would be incorrect reporting of the original models, particularly for degrees of freedom (i.e., failure to report a negative error variance being fixed to some positive value) and parameter estimates (i.e., failure to report correct correlation coefficients, factor loadings, or variances).

**Summary and Implications**

Regarding Research Question 1, information necessary for reanalysis was obtained from only four out of the 33 authors (12%) we contacted. Such information was not provided in the articles primarily because doing so was not requested by the editor(s) or reviewers.

Thus, the success rate for getting the data was lower than but not very different from previous studies where success rate for acquiring psychological data from the authors was 24% (Wolins, 1962), 26% (Wickerts et al., 2006), and 38% (Craig & Reese, 1973). Since the information we requested (i.e., the variance/covariance or correlation matrices [with means
and $SD$s)) was considerably manageable compared with the raw data requested in these previous studies, we initially expected to be more successful retrieving data, only to perform worse. Many of the journals in second language testing and learning follow the *Publication Manual of the American Psychological Association*, in which data sharing is encouraged. Nevertheless, our mostly failed attempt to obtain data from the authors suggests in reality a very limited availability of data from the authors, regardless of disciplines and types of data requested. Reasons for unwillingness to share data for replication of structural equation models are not clear because we did not send a second e-mail requesting the data or an explanation of why it was not forthcoming. Future research is needed in this area by sending out a questionnaire to all researchers asking why they are hesitant to cooperate.

Second, 124 of the 360 models we gathered were reanalyzed for Research Question 2. The success rates of replication of structural equation models were 89% in the preliminary analysis, 87-100% for fit indices, and 94% for parameter estimates. These highly successful replication rates of structural equation models make researchers confident that they can believe what they read and that they can base their new study on earlier work.

Three implications are discussed. First, as seen from the difficulty of retrieving data from the authors, we suggest that journal editors and reviewers require reporting of (a) the variance/covariance matrix and/or (b) the correlation matrix with means and $SD$s for any SEM analysis at the time of paper submission. This enables editors and reviewers to conduct
model replication. Despite space limitations, such matrices should be provided at the article reviewing stage. Second, we suggest that data for replication are archived at journal or authors’ homepages, as this (a) precludes errors in manually inputting data from articles, (b) avoids paper length constraints, and (c) reduces the situation where renowned researchers contacting the authors of the original studies are more likely to receive responses to inquiries than less renowned researchers are. Third, reanalysis of structural equation models should be encouraged more in testing and assessment, especially in second language testing and learning. We suggest that before undertaking new studies, researchers reanalyze existing structural equation models. If the information necessary for reanalysis is not reported, they should contact the authors. As such interactions increase, the authors may recognize the value of reporting such information in the first place.

Acknowledgement

We thank the four authors for kindly sharing their data, Yuko Hoshino, Stephen G. Sireci, and the anonymous reviewers for their valuable comments on earlier versions of this paper. This research was supported by Grants-in-Aid for Scientific Research (No. 20820020) provided by Japan Society for the Promotion of Science.
References


317–340.


measuring anxiety at different stages of the foreign language learning process: The input anxiety scale, the processing anxiety scale, and the output anxiety scale. *Language Learning, 50*, 87–117.


## Absolute Value Differences in Model Fit Indices and in Parameters Between Original and Replicated Models

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Median</th>
<th>Outlier n</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>105</td>
<td>121.800</td>
<td>696.471</td>
<td>0.000</td>
<td>6932.187</td>
<td>9.267</td>
<td>90.263</td>
<td>0.950</td>
<td>2 (2%)</td>
</tr>
<tr>
<td>$p$</td>
<td>78</td>
<td>0.071</td>
<td>0.176</td>
<td>0.000</td>
<td>0.920</td>
<td>3.139</td>
<td>10.189</td>
<td>0.000</td>
<td>5 (6%)</td>
</tr>
<tr>
<td>$df$</td>
<td>104</td>
<td>1.644</td>
<td>6.903</td>
<td>0.000</td>
<td>66.000</td>
<td>8.140</td>
<td>74.836</td>
<td>0.000</td>
<td>1 (0.01%)</td>
</tr>
<tr>
<td>$\chi^2/df$</td>
<td>23</td>
<td>0.308</td>
<td>0.707</td>
<td>0.000</td>
<td>3.009</td>
<td>3.174</td>
<td>10.412</td>
<td>0.010</td>
<td>2 (9%)</td>
</tr>
<tr>
<td>NFI</td>
<td>24</td>
<td>0.036</td>
<td>0.052</td>
<td>0.000</td>
<td>0.217</td>
<td>2.214</td>
<td>5.670</td>
<td>0.015</td>
<td>1 (4%)</td>
</tr>
<tr>
<td>NNFI</td>
<td>41</td>
<td>0.013</td>
<td>0.026</td>
<td>0.000</td>
<td>0.136</td>
<td>3.355</td>
<td>13.486</td>
<td>0.000</td>
<td>2 (5%)</td>
</tr>
<tr>
<td>CFI</td>
<td>64</td>
<td>0.012</td>
<td>0.029</td>
<td>0.000</td>
<td>0.155</td>
<td>3.646</td>
<td>13.764</td>
<td>0.000</td>
<td>4 (6%)</td>
</tr>
<tr>
<td>GFI</td>
<td>31</td>
<td>0.005</td>
<td>0.018</td>
<td>0.000</td>
<td>0.102</td>
<td>5.137</td>
<td>27.451</td>
<td>0.000</td>
<td>1 (3%)</td>
</tr>
<tr>
<td>AGFI</td>
<td>8</td>
<td>0.021</td>
<td>0.060</td>
<td>0.000</td>
<td>0.169</td>
<td>2.828</td>
<td>7.997</td>
<td>0.000</td>
<td>1 (13%)</td>
</tr>
<tr>
<td>RMSR</td>
<td>9</td>
<td>0.004</td>
<td>0.009</td>
<td>0.000</td>
<td>0.027</td>
<td>2.671</td>
<td>7.435</td>
<td>0.000</td>
<td>1 (0.1%)</td>
</tr>
<tr>
<td>RMSEA</td>
<td>53</td>
<td>0.010</td>
<td>0.018</td>
<td>0.000</td>
<td>0.103</td>
<td>3.673</td>
<td>15.429</td>
<td>0.000</td>
<td>3 (6%)</td>
</tr>
<tr>
<td>upper</td>
<td>4</td>
<td>0.015</td>
<td>0.017</td>
<td>0.000</td>
<td>0.040</td>
<td>1.540</td>
<td>2.889</td>
<td>0.010</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>lower</td>
<td>4</td>
<td>0.015</td>
<td>0.017</td>
<td>0.000</td>
<td>0.040</td>
<td>1.540</td>
<td>2.889</td>
<td>0.010</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>AIC</td>
<td>4</td>
<td>83.063</td>
<td>104.090</td>
<td>25.680</td>
<td>238.870</td>
<td>1.975</td>
<td>3.913</td>
<td>33.850</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>CAIC</td>
<td>4</td>
<td>81.980</td>
<td>104.886</td>
<td>22.750</td>
<td>238.870</td>
<td>1.967</td>
<td>3.884</td>
<td>33.150</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

### Parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Median</th>
<th>Outlier n</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>35</td>
<td>0.249</td>
<td>0.404</td>
<td>0.000</td>
<td>1.780</td>
<td>2.284</td>
<td>5.529</td>
<td>0.040</td>
<td>2 (6%)</td>
</tr>
</tbody>
</table>

**Note.** $N$ = the number of models with the fit indices or parameters reported. Although the total number of models reanalyzed was 105, not all the models reported fit indices and/or parameters. For example, chi-squares were reported in all of the studies ($N = 105$), and these 105 chi-square values originally reported were compared with 105 chi-square values replicated in the current study. On the other hand, $p$ values were reported in 78 out of the 105 studies. Outlier $n$ = the number of models whose absolute value differences in model fit indices and parameter estimates between the original and replicated models lies out of the mean $+$/− 2SD (thus judged to be outliers, or unsuccessful replication). *All parameter values are standardized except for one unstandardized value.*