In’nami, Y., & Koizumi, R. (2011). Factor structure of the revised TOEIC test: A multiple-sample analysis. *Language Testing, 29,* 131-152.

Abstract

- This study examined the factor structure of the listening and reading section of the revised TOEIC test.
- Four models (higher-order, correlated, uncorrelated, and unitary) were hypothesized and tested.
- The results suggested the correlated model fit the data best.

1. Literature review

- Though factor structure of tests would contribute to understanding how test scores relate to the constructs being measured, few research examined the factor structure of TOEIC test. This study aims to narrow this gap.

1.1 TOEIC test

- The TOEIC test were revised in 2006.
- The revisions of TOEIC test are as follows (Schedl, 2010):

(1) listening & reading	Passages become longer.
(2) listening section	The listening materials are recorded by voice actors of various English speakers.
(3)	The picture description task has fewer questions, while the short talk task has larger questions.
(4) reading section	The error recognition task has been replaced with a task requiring the examinee to fill in the blanks in complete passages.
(5)	The double-passage questions are added, which the examinee compares the two related passages.

1.2 Validation studies on the TOEIC test

- There are far fewer validation studies on the TOEIC test than the TOEFL or IELTS. These few studies can be divided into 3 categories based on their purpose: reliability and score distribution, variables related to score gain, and relationships with other measurement scales.
- Only Willson’s study (2000) examined the factor structure of the ability measured in the old version of TOEIC test targeting Japanese and Korean. With expository factor analysis, he found unidimensionality for the listening section but bidemensionality for the reading section. However, there is no study on the factor structure of the revised TOEIC test.
- As for the factor structure of the revised TOEIC test, the examinees receive a single total score along with separate scores for the listening and reading sections. The use of single total score assumes that a single high-order or hierarchical factor underlies performance on both the listening and reading, whereas the use of separate scores assumes that distinctive factors of listening and reading skills are involved.
- The authors hypothesize that the factor structure of the revised TOEIC test is hierarchical, where a high-order factor of the receptive skill influences listening and reading skills.

1.3 Structure of Language ability

■ The inconsistent relationships between listening and reading skills found across studies lead us to hypothesize as follows.

(a) Listening and reading are inseparable;

Oller (1983) analyzed the students' response on the placement test with Principle component analysis and reported L2 language ability was a single trait (though this hypothesis was refuted).

(b) Listening and reading are separable and uncorrelated; Wilson (2000)

(c) Listening and reading are separable but closely correlated;

Bae and Bachman (1998) used the exploratory factor analysis, and concluded as described above.

(d) Listening and reading are hierarchically structured;

Song's (2008) study with the confirmatory factor analysis revealed one common comprehension factor influenced 3 subskills.

1.4 Cross-validation with multiple-sample analysis

■ In addition, the generalizability of a factor structure can be examine by using a cross-validation method. If the same model fit the data well across samples, this would suggest the generalizability of the model.

Research questions

RQ1: Does the higher-order model assumed in the revised TOEIC test fit the data better than the correlated, uncorrelated, and unitary models?

RQ2: Is the factor structure of the revised TOEIC test generalizable across samples?

2. Method

2.1 Data

■ Data were obtained from 569 English learners in Japanese university. Their nationalities were Japan, South Korea, and other Asian country. Notice that this population was not quite representative of the TOEIC.

■ The data were derived from the TOEIC IP test.

2.2 Analyses

■ The data were provided by the TOEIC in the form of the percentage of correct scores.

■ The 4 listening subskills:

(a, b) Infer gist, purpose, and basic context based on explicit information in short and extended spoken texts

(c, d) understand details in short and extended spoken texts

■ The 5 reading subskills:

(a) Make inferences in written texts

(b) Locate and understand specific information in written texts

(c) Connect information across multiple sentences in a single written text and across texts

(d) Understand vocabulary in written texts, and (e) Understand grammar in written texts

- In order to conduct cross-validation analyses, the data were randomly split into two groups (each *n* ≒ 285).
- Confirmatory factor analysis was used to investigate the factor structure of TOEIC Test. The following 4 models were tested: a higher-order trait model, a correlated trait model, an uncorrelated trait model, and a unitary trait model.
- Maximum likelihood method was used to estimate model parameters.
- Model fit was evaluated by a non-significant chi-square; CFI, NFI, and TLI of .90 or above; RMSEA of 0.05 or below; SRMR of 0.08 or below; lower values of AIC and CAIC.
- As preliminary analyses, the univariate and multivariate normality were confirmed.

3. Results

3.2 Testing of the four models with each samples

- Higher-order model fits the data well, but the chi-square statistics was significant.
- Correlated model fits as well as the higher-order model, and was substantively interpretable. → O
- Unitary model was statistically less favorable than the correlated model.
- Uncorrelated model showed poor fit across the samples.

3.3 Multiple-sample analysis

- Cross-validation of the correlated model was conducted as follows.
- Model is gradually restricted by adding constraints to examine the extent to which is across samples.

(Model 1) configural invariance

(2) invariance of factor loadings

(3) invariance of both the factor loadings and the measurement error variances

(4) invariance of the factor loadings, measurement error variances, and factor variances

(5) invariance of the factor loadings, measurement error variances, factor variances, and factor covariances

→ All models fit the data well. Accordingly, they were compared using chi-square tests and CFI to determine which model was more stable across the two samples.

- Based on CFI, since the most stringently tested Model 5 fit the data well across the two samples, Model 5 was adopted.

4. Discussion and conclusion

- Answer to RQ1:

  The correlated model was selected as the best model for both samples for following two reasons:

(1) Both the correlated model and the higher-order model fit the data better than the uncorrelated and unitary model.

(2) Selecting the higher-order model as the final model was not appropriate because the variance in the model needed to be fixed to obtain model identification.

- Answer to RQ2:

  The correlated factor structure identified for the revised TOEIC test could be generalized across samples.

■ The result of correlated factor structure for TOEIC test was consistent with some previous studies on the factor structure of L2 ability (e.g., Bachman & Palmer, 1981).

■ However, this result was not consistent with the study of the old version of the TOEIC test (Wilson, 2000).

→ The old version of TOEIC test includes unidimensionality for listening comprehension, and bidimentsionality for reading comprehension (Wilson, 2000). There are three possible explanations for these differences.

(1) Differences in the content

■ The revised TOEIC test does not include the error recognition task as seen in the old TOEIC test. Wilson regarded the error recognition task as a separate factor apart from a general reading comprehension factor.

→ The deletion of the error recognition task in the revised TOEIC test might have produced the unidimentionality of the reading skill.

(2) Difference of analytical methods

■ Wilson used exploratory factor analysis with varimax rotation, because his purpose was to find out the factor. That method is utilized to extract orthogonal/ uncorrelated factors.

■ In the current study, the models to examine were decided beforehand. Thus, confirmatory factor analysis with promax rotation was used, which extract oblique/ correlated factors.

(3) Difference of parcels for factor analysis

■ Wilson made the large number of parcels from the raw data, whereas the current study made the parcels from the percentage of correct score.

■ Having a large number of parcels as Wilson had would make it difficult for a model to fit the data well although it could be a good model indeed.

5. Implications and limitations

■ Two main implications:

(1) Relatively high correlation between the listening and reading factors suggested the distinct but highly nature of these two skills and support single score reporting, thus the current results provide empirical support for the revised TOEIC test.

(2) Usefulness of testing for the invariance of a factor structure

■ More cross-validation studies with invariance tests can be conducted, since datasets from large-scale tests are often large enough to yield satisfactory sample sizes for multiple groups.

■ Three limitations:

(1) Unavailability of the item-level data could preclude appropriate analysis.

(2) The current study finding is limited to the Japanese sample and is not generalizable to the TOEIC test-taking population.

(3) The TOEIC-IP data used in this study were drawn from one of the several forms of the test. The another study is needed to see whether the correlated factor structure will be supported in the other forms of the revised TOEIC tests.

前回の補足

■　因子分析の前提 (三浦, 2008)

(1) ある程度の標本数を確保すること

□　絶対的な基準はなく、項目数・回答の方法 (3 件法など) ・抽出する因子の数によって必要となる標本数が異なる。

□　安定した結果を得るためには、少なくとも 100 人程度の被験者が必要である。変数 (項目) が増えれば、さらに多くの被験者が必要で、一般には項目の 4 倍くらいの被験者が必要だと言われている。(狩野・三浦, 2002)

➔ 今回の研究に関しては、下位尺度が 9 項目 (リスニング 4 項目・リーディング 5 項目) あるので、協力者が約 285 名であることを考えると、サンプルサイズが足りないと言える。

□　因子内の項目数が増えれば、必要なサンプル数は減る。

(2) 観測変数は間隔尺度以上であること

■　因子分析の標本数が足りない場合の修正法

➔ 参考文献には書かれていなかったので、修正方法として一般的なものはないのではないか。
ただし、因子分析のサンプルサイズを定めた統一的な記述もないので、研究を通して基準が一貫していれば良いのでは?

■　因子分析はアンケートでなくてもできるのか?

□　因子分析は、テスト結果の項目そのものを扱って、学習者の能力を検証するのにも使用される。

□　アンケートは順序尺度であることが多く (よくあてはまる⇔まったくあてはまらない) 、これは厳密には間隔尺度のデータとして処理するのは適切でない。ただし、5 件法以上の順序尺度であれば間隔尺度以上とみなしても結果に大きな影響は出ないとされることから、因子分析が使用される。(山上・倉智, 2003)

→ アンケートは因子分析の応用であり、テスト得点の方が因子分析には適切?

■　Parcel factor analysis

いくつかの項目をまとめて下位尺度とする方法は小包化（parceling）と呼ばれ、下位尺度を観測変数とする探索的因子分析法を小包因子分析（parcelfactor analysis）とも言う。小包化したほうがより適切な解を推定する可能性が高まる。

< 参考文献 >

狩野裕・三浦麻子. (2002). 『グラフィカル多変量解析』. 京都: 現代数学者.

三浦省吾 (監修). (2008). 『英語教師のための教育データ分析入門　―授業が変わるテスト・評価・研究―』. 東京: 大修館書店.

山上暁・倉智佐一. (2003). 『要説　心理統計法』. 京都: 北大路書房.