Journal of Language Teaching, *3*(5), 34-38, 2023
https://doi.org/10.54475/jlt.2023.014

Review

Review on the validity of China's Standards of English Language Ability

Lu Wang, Meihua Chen

School of Foreign Languages, Southeast University, Nanjing, China

Received: April 17, 2023 / Accepted: April 28, 2023 / Published Online: May 1, 2023

Abstract

The English and Chinese versions of China's Standards of English Language Ability (CSE) were released in 2018. The appearance of CSE helps to solve the problems of different standards of English exams in China, the separation of teaching and assessment objectives, and the incoherence of teaching objectives at various stages. Before, during, and after the development of the CSE, many scholars have discussed the construction of the scale from theory to practice and contributed to the realization of same standardization for English testing in China. This paper aims to review the relevant studies in order to provide insights and suggestions for the future research and application of the CSE from three aspects: 1) introducing the two major theoretical frameworks for validating language scales in China; 2) reviewing the studies on the validity of the CSE in general; and 3) reviewing the empirical studies on the validity of the sub-scales in the CSE, including listening, speaking, reading, writing, interpreting, translation, pragmatic competence. However, there is a lack of studies on the aspect of organizational competence.

Keywords China's Standards of English Language Ability, validity, language scales

1. Introduction

The main development of the CSE was completed at the end of 2016, and on February 12, 2018, it was officially released by the Ministry of Education and the State Language and Literature Working Committee and was officially implemented on June 1, 2018 (State Language Commission, Ministry of Education & State Language Affairs Commission, 2018). CSE is oriented to language use and divides learners' English proficiency into three stages, namely, basic, advanced, and proficient level. The release of the CSE helps solve the problems of different standards of English exams, separation of teaching and assessment objectives, and incoherence of teaching objectives at each stage, and achieve a one-stop English teaching process and mutual recognition of multiple learning outcomes.

CSE includes a language proficiency matrix, as well as a proficiency matrix for listening comprehension, reading comprehension, oral expression, written expression, organizational competence, pragmatic ability, interpretation, and translation ability, etc. (Ministry of Education & State Language Affairs Commission, 2018). Since its release, it has been gradually applied in language learning, teaching, and testing. Many researchers showed great concern on the validity of CSE and there have been many studies from different perspectives. Currently, the research on the validity of CSE mainly focuses on the overall validity of CSE, as well as on the validity of each scale, including listening, speaking, reading, writing, interpreting and translation scales, etc.

2. Literature Review

2.1. Two major theoretical frameworks for validating language scales

Language scales can measure the language ability of participants. Therefore, the validity of a scale can be defined as the extent to which a test measure what it is supposed to measure (Chapelle, 1999). Before the establishment of CSE, there have been some pivotal studies in the field of the validity of language scales (Li, 2020). Two major theoretical frameworks for validating language scales were proposed in the following two studies.

Zhu (2016) defined the basic content of the research on the validity of language scales and provided a theoretical framework for validating

National English Proficiency Scale of China (NEPS) in his study, which can be generally referred to as the "social and educational cognitive model" (Zhu, 2016, p.9). In this framework, the validity of a language scale is "the extent to which the scale measures the target language ability constructs" (Zhu, 2016, p.3). On the one hand, the author discussed the connotation and interconnection of construct validity research and fairness validity research from the perspectives of science and ethics. On the other hand, the author discussed the importance and essentiality of teaching backwash validity and social impact validity from the perspectives of English education and social life. To summarize, the NEPS should be scientific, fair, valid for relevant decisions, and should have a positive impact on English language teaching and social life. Moreover, this framework specifies various methods of evidence collection, including questionnaires, field surveys, interviews, psychological experiments, statistical methods and big data analysis, etc.

In the next year, Fang & Yang (2017) proposed a validation framework for validating the scale, including four types of validity, namely, construct validity, content validity, criterion validity, and use validity. The framework considers that construct validity and content validity belong to the internal validity of the scale, while the validity of criterion validity and use validity belong to external validity of the scale. Internal validity is the first priority, which determines the external validity to a large extent. There is no specific discussion over the research methods of validity in this framework. However, two basic requirements of validating scales were proposed, one is that scientific and operational validity be given equal importance, and the other is that valid experiments and surveys be conducted. Moreover, the authors proposed that the construct validity of CSE is mainly manifested in the following three aspects, 1) the adaptability of the scale to the specific language teaching and testing social environment; 2) the rationality of the scale's intended goals and its' usage; and 3) the scientific validity and feasibility of the theoretical rationale, ideas and methods used to develop the scale. Constructs are mental processes or characteristics that explain differences in the behavior of individuals or groups, and construct validity refers to the extent to which a measure measures the construct to be measured (Strauss & Smith, 2009). In the construction of the scale, construct validity is the extent to which the scale reflects the competencies to be included in the scale (Luiz et al., 2001), and it is related to the state of language education and language proficiency theory in a given social context.

The two frameworks have different conceptual names, crossover between validity categories, and slightly different categories. Zhu's framework emphasizes the primacy of decision validity, while the Fang and Yang's framework puts emphasis on the primacy of internal validity (construct and content validity). The strength of the Zhu's framework is that it highlights the importance of fairness and consequences (teaching backwash validity and social impact validity), and the strength of the Fang and Yang's framework lies in the clearer definition of construct and content validity, and it is more operational.

2.2. Research on the overall validity of CSE

After the release of CSE in 2018, some studies began to focus on the validity of CSE. The following two are the most influential articles which analyze the validity of the CSE in general.

Liu (2021) tested the construct validity, fairness validity, and procedural validity of CSE based on the Assessment User Argument (AUA) validation model, and this study revealed strong evidence in support of the overall validity of CSE. Fairness validity refers to the degree of fairness of the examination, that is, all parties related to the examination should be fair and impartial at any stage of the examination, from the design of the examination to the use of the results, and there should be no improper factors such as non-examination-related conceptions and misuse of examination results. For the scale like CSE, there should not be any bias on the gender, race, religion or culture in the description when conducting the Differential Item Functioning (DIF) analysis to test the fairness validity (Zhu, 2016). Procedural validity refers to the appropriateness of procedures and the quality of their implementation (Kane, 1994), which includes clarity, operability, and reasonableness of the procedures (Pant et al., 2009). The procedural validity of CSE is to test whether the procedures adopted in the development of the scale are realistic and whether the design of the steps is appropriate and scientific, and whether each step is well executed (Papageorgiou & Tannenbaum, 2016). In Liu's (2021) study, there were altogether 130 thousand participants and 30 thousand English teachers involved in the validation. Due to the specialty of the CSE, a validation model was constructed in this study based on Toulmin's (2003) validity argument theory and Bachman & Palmer's (2010) AUA theoretical model. Results showed good construct validity, fairness validity and procedural validity of the CSE with quantitative and qualitive data.

In order to validate the self-assessment grids of CSE, Zhou (2021) adopted the validity framework of Chapelle et al. (2011) to construct an IUA framework, which consists of four types of reasoning: scoring, generalization, interpretation, and extrapolation. This study used statistical methods to test the five assumptions proposed in the framework. The study indicated that the scale consists of descriptors of different levels of difficulty, which can reliably distinguish students of different English levels. The difficulty level of the descriptors at each level in this study increased as the level increased, and the difficulty level of the descriptors basically matched the language proficiency levels specified in the scale, supporting the generalized inference of the self-assessment scale. The correlation between the self-assessment results and the standardized test results, although weak, was significant and largely consistent with the results of existing studies, thus largely supporting extrapolative inference. In general, multiple evidence suggest that the self-assessment scales have good validity.

2.3. Research on the validity of the sub-scales in CSE

Liu & Han (2018) constructed a theoretical framework for the application-oriented language proficiency scale, which classified the various competencies in the scale into listening, speaking, reading, writing, interpreting, translation, pragmatics, and organization based on the actual situation of language learners and users' proficiency levels and the degree of social needs. The following passage will also provide an overview of the empirical research on the sub-scales in these areas.

He & Chen (2017) validated listening ability subscale of the CSE in terms of ability conceptualization, rating, and usage of the scale. They defined construct validity as "the extent to which the descriptive and parametric frameworks of the scale reflect ability constructs". From their descriptions, it is clear that the scale developers set up the listening ability model based on the actual needs of English teaching and testing in China and the latest research results of listening comprehension at home and abroad. The authors also proposed the parameter framework for the descriptors accordingly, which was repeatedly validated by relevant experts. In addition, the interview data of teachers and students are also evidence of the construct validity of the descriptors. To ensure the validity of the scale rating, the scale developers used a combination of qualitative and quantitative methods. This study emphasized the importance of post validity evidence for CSE use, arguing that applied research in different domains is an important source of validity evidence.

In order to validate the oral ability subscale of CSE, Wang (2020) adopted the text-mining approach to compare the similarities and differences between CSE and Common European Framework of Reference (CEFR) in terms of the following three aspects, namely, themes, co-occurrence network and distinguishing features in each level. In this study, a text mining software was used to analyze the content of all descriptors of CSE and CEFR oral communication activities. Comparing the typical characteristics of the descriptors at different levels of the two scales, a high degree of similarity was found between the two verbal expression descriptors. For example, both CSE Levels 1 and 2 and CEFR Level A1 describe verbal expression using simple language; both CSE Level 5 and CEFR Level B1 emphasize personal opinions on relevant issues in verbal expression; CSE Levels 8 and 9 and CEFR Levels C1 and C2 both describe the use of complex language for effective communication and exchange in the professional domain. However, there are also differences between the two, for example, the CSE oral ability subscales have close semantic relationships and are clustered together, especially at CSE levels 1 and 2, 8 and 9, whereas the six CEFR levels are relatively dispersed and the semantic distance between levels is relatively far. Nevertheless, the general results of the study showed that the two scales had greater similarities than differences, which indicates to some extent that the CSE oral ability subscale has a high validity. The findings also suggest that the descriptors in part of the adjacent levels are not clear-cut.

Zhou (2021) verified the validity of the reading strategy descriptors of the CSE at the higher education level from the perspective of the Rasch measurement model. The Rasch measurement model was applied to verify the validity of the descriptors as follows. First of all, the author compared the actual ranking of the topics from easy to difficult with the expected ranking. The expected ranking of topic difficulty can be based on expert judgment, existing research, or a combination of both. Then, she compared the spacing of topics with the expected spacing, and examined the Differential Item Functioning (DIF) of the topics. If a topic exhibits a DIF, it means that the traits measured by the topic are defined differently for different groups. The participants in this study included 30,772 questionnaire takers and 12 interviewees. This study showed that the reading strategy descriptors fit well, and the overall difficulty ranking of the descriptors was consistent with the expert's judgment. The overall difficulty ranking of the descriptors is consistent with the experts' judgment, and the level classification is basically reasonable. However, there are still a few descriptors whose difficulty ranking is different from the experts' predicted difficulty, and the number of levels differed slightly from the experts' predictions. The authors identified the problematic descriptors based on the validity verification, and ensured the clarity and de-jargonization of the descriptors through deletion, so that the level representation of the descriptors was optimized.

CSE writing scales consist of two subscales, namely, written expression ability and written expression strategies. The validation of the CSE writing scales included expert judgment, two graded validations, and in-depth interviews. The content validity of the descriptors and the rationality of the descriptor classification were examined by expert judgment. In-depth interviews revealed the factors that influence the inconsistency of some descriptor classifications with expectations. Deng, Deng & Zhang (2021) validated the writing scales of CSE, focusing on the content, categorization and grading of descriptors. The results in this study showed that the writing scale descriptors were comprehensive and typical, the categories were reasonable, the descriptors had great goodness-of-fit, the overall difficulty level was basically consistent with expert judgment, and the level division was basically reasonable. Based on the validation results, the writing project team processed the descriptors to ensure that the descriptors were comprehensive, typical, and relevant in content, correct and non-crossing categories, and monotonically increasing difficulty levels with good differentiation. The validation of the descriptors in the development stage of the scale can ensure the practicality of the scale and guarantee its full implementation. This study can provide a reference for the validation of the CSE writing application scale, and can help accelerate the construction of English writing assessment standards in China.

Xu, Yang & Mu (2019) pointed out that the validation of the interpreting ability descriptors consisted of two graded validation processes. The first one was conducted by a quantitative method using a large-scale data survey to determine the level of descriptors by means of a descriptor questionnaire for the relevant population groups, i.e., learners, users and teachers of the corresponding levels. In the second validation process, a qualitative approach was used to conduct focus group interviews with users of the interpreting scale to explore the appropriateness and usefulness of the descriptors for interpreting ability. The results of large-scale quantitative cross-validation showed that the descriptors of the interpretation scale had a moderate goodness-of-fit. However, there are still some unfitting descriptors, low differentiation descriptors and a large number of difficulty parameters that do not match the original level. Therefore, the first validation provided data for further adjustment and modification of the descriptors. The second graded validation showed that some of the proficiency descriptors in the interpretation scale were repeated or similar descriptors. In response to the inconsistency, incomprehensibility, ambiguity and repetition of the descriptors, the descriptors were revised one by one after the second graded validation.

Lv & Ren (2022) adopted Rasch's rating scale model to examine the validity of the translation ability scale of CSE. A self-assessment survey was conducted to collect data from students and practitioners on the 33 descriptors of the scale. The study found that the RSM model can effectively estimate the difficulty and differentiation of the descriptors, which can help to screen out the poor-quality descriptors; the overall reliability of the descriptors is high, and they have good conceptual validity; the scale can distinguish between different levels of participants. These findings provide necessary data support for the future application of the scale in the teaching and evaluation of translation. However, the study is limited in the sample size and the lack of qualitative data analysis.

The pragmatic competence scale of the CSE is based on two dimensions, namely, language comprehension ability and language expression ability, and the scale classifies learners' language proficiency into nine levels from low to high, and describes the performance characteristics of each proficiency level to provide a guide for learners to self-assess their language proficiency. Sun & Fu (2021) verified the validity of the pragmatic competence scale of the CSE from the perspective of self-assessment by Multi-faceted Rasch Model based on AUA. In this study, the validity of the pragmatic competence scale was interpreted in terms of the degrees of agreement and discrimination. The former was mainly examined in terms of the consistency in the severity of ratings among learners, which was reflected in the goodness-of-fit of descriptors and rating scales; the latter was judged mainly with reference to two indicators, namely, the separation coefficient and the reliability of the separation coefficient.

3. Conclusion

Since the release of the CSE, an increasing number of researchers and scholars began to pay attention to the validity of the scale, whether the general validity of the scale or the validity of sub-scales in the CSE. However, to the best of our knowledge, there is a lack of research in the field of the organizational competence in the CSE. Besides, the number of the published paper is still not vey enough in the other sub-scales as well as the validity of the CSE in general. Therefore, much more attention could be paid to the study of the validation of CSE in the future.

Lu Wang is an M.A. candidate in Applied Linguistics at Southeast University. She obtained a B.A. in English literature at JiangNan University, China. She has published research articles in journals such as Technology Innovation and Application. She worked as a TA at Southeast University. Her research interests include Language Policy and Planning, self-regulated language learning, and language testing.
Email: luwanggrace98@gmail.com

Meihua Chen (Corresponding Author) is Professor of Linguistics in the School of Foreign Languages, Southeast University. She is a PhD supervisor. Her research has appeared in a broad array of prestigious journals such as Frontiers in Psychology, Foreign Language Teaching and Research, Foreign Languages in China, Foreign Language Education, Foreign Languages and Their Teaching, Technology Enhanced Foreign Languages. Her main research interests are in the field of Applied Linguistics, English Language Education, Language Policy and Planning.
Email: meihuachen123@126.com

References

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254-272. https://doi.org/10.1017/S0267190599190135

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2011). *Building a validity argument for the Test of English as a Foreign Language TM.* Routledge.

Deng, H., Deng, J. & Zhang, W. X. (2021). Validation of the writing scales of the China's Standards of English Language Ability. *Foreign Language*

World (05), 66-74.

Fang, X. J. & Yang, H. Z. (2017). Validity and Validation of Language Proficiency Scales. *Journal of Foreign Languages* (04), 2-14.

He, L. Z. & Chen, D. J. (2017). Investigating the inner structure of China's Standards of English: Descriptive scheme and salient features of listening ability descriptors. *Foreign Language World* (4), 12-19.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*(3), 425-461. https://doi.org/10.3102/00346543064003425

Li, Q. H. & Kong, S. (2020). Research on Validity and Validation of Language Proficiency Scales: A Review. *Journal of Beijing International Studies University* (05), 32-45.

Liu, J. D. (2021). Validating China's Standards of English Language Ability. *Modern Foreign Languages* (1): 86-100.

Liu, J. D. & Han, B. C. (2018). Theoretical considerations for developing use-oriented China's Standards of English. *Modern Foreign Languages* (1): 78-90.

Luiz, D. M., Foxcroft, C. D., & Stewart, R. (2001). The construct validity of the Griffiths Scales of Mental Development. *Child: Care, Health and Development*, *27*(1), 73-83. https://doi.org/10.1046/j.1365-2214.2001.00158.x

Lv, X. X. & Ren, W. The validation of translation self-assessment scale from China's Standards of English based on the Rasch Model. *Foreign Language Education*, (01), 57-61.

Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P., & Köller, O. (2009). Validity issues in standard-setting studies. *Studies in Educational Evaluation*, *35*(2-3), 95-101. https://doi.org/10.1016/j.stueduc.2009.10.008

Papageorgiou, S., & Tannenbaum, R. J. (2016). Situating standard setting within argument-based validity. *Language Assessment Quarterly*, *13*(2), 109-123. https://doi.org/10.1080/15434303.2016.1149857

State Language Commission, Ministry of Education & State Language Affairs Commission. (2018). *China's Standards of English Language Ability*. Higher Education Press.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, 5, 1-25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Sun, L. & Fu, X. H. (2021). Research on the self-assessment validity of pragmatic competence scale based on Multi-faceted Rasch Model. *Journal of Xi'an International Studies University* (02), 52-57.

Toulmin, S. E. (2003). *The Uses of Argument (Updated edition)*. Cambridge University Press.

Wang, H. (2020). Research on the validity of speaking scale in China's Standards of English: Based on the text-mining approach. *Journal of Xi'an International Studies University* (2), 69-74.

Xu, Y., Yang, Y. & Mu, L. (2019). Validating descriptor levels of the Interpreting Scale of the China's Standards of English Language Ability. *Foreign Language World*, (4), 24-31, 66.

Zhou, Y. Q. Validating the Self-assessment Grids of China's Standards of English Language Ability. *Modern Foreign Languages* (01), 101-112.

Zhou, Y. Q. Validating descriptors of the reading strategic competence subscale of CSE: Rasch measurement model perspective. *Foreign Language World*, (1), 79-87

Zhu, Z. C. (2016). A validation framework for the National English Proficiency Scale of China. *Journal of China Examinations*, (08), 3-13.