〈研究ノート〉

Defining Validity for a University EFL Placement Test

Robert J.S. ROWLAND

Abstract

This paper is a collection of research notes regarding how to define validity for an EFL placement exam implemented at a low-level university in Japan. Level placement tests are common in university EFL programs, but there is little research into their validity. In this paper, the literature was examined to determine the necessity of level division for English classes at Japanese universities, different methods for level division, and validity considerations for using a standardized test for level division. This preliminary research will inform future validity studies into the EFL placement exam at a small university in Japan.

Key words: Placement test, University EFL, Validity, Standardized tests

Many universities offer English language classes in multiple ability levels. Students are divided into levels before or upon entry to an English language program using a level placement test. Although use of a placement test to divide students into levels is common practice in a variety of language learning contexts (Richards & Schmidt, 2002), research into the validity of level placement test design or implementation is lacking (Papageorgiou & Cho, 2014). Existing research into placement testing has examined the validity of use in English for Academic Purposes (EAP) classes in tertiary education (Fulcher, 1997), and to a lesser extent, in English as a Second Language (ESL) classes (Papageorgiou & Cho, 2014). Little research exists, however, into their use in the English as a Foreign Language (EFL) context, especially in Japan. The small, liberal arts university at which the researcher taught has used level placement tests in the English language education program for years. The tests used have changed overtime, but there has been little academic examination into the validity of these tests. As a prelude to a formal validity analysis, this paper examines the literature determine what must be considered when implementing and assessing the validity of a level placement test at a university in Japan to better assess and improve the placement procedure.

人文学部・欧米文化学科

論文受理日 2022 年 11 月 21 日

The Necessity of Levels

Level placement is so common in Japanese university English programs that its necessity is rarely questioned. To fully assess the validity of a level test, however, one must first be able to argue the consequential validity of the test, that is, whether the purpose for which it is used is indeed a valid one. Given that purpose of a placement test is to place students into groups based on specific criteria, first we must define was it meant by "place." Green (2012) stated that there are two different ways to think about the benefits of the concept of student "placement." The first is to focus on the placement of all students on a spectrum of ability, to better understand how to divide students into meaningful ability groups with their peers. The second is to place students on a spectrum of difficulty with regards to course materials to ensure students interact with curricula that best fits their ability. While level placement seems an intuitively appropriate practice in any educational institution with a rage of student knowledge and abilities, there is disagreement in the research about its efficacy.

Supporters of ability grouping cite a range of positive effects on both learners and teachers. Such benefits include ease of class planning and organization (Brown, 1995), ease of on-the-fly curriculum adjustments to better fit the needs of a specific group of students (Kim, 2012), and lower learner anxiety levels in the classroom (Lou & Tsai, 2002). It has also long been argued in the language learning literature that learners are best equipped to learn concepts that are just outside their current realm of knowledge (Krashen, 1985; Lightbown & Spada, 2013). Thus, grouping students of similar ability levels together may maximize the efficiency of instruction by facilitating curriculum design and implementation targeting more specific learner goals.

There are also arguments against grouping students by ability level. From a student perspective, placement into a given level may have a negative effect on student identity. Ames (1992) proposed that learners placed in higher level classes could develop an imposter syndrome, a feeling that they do not belong in the higher level. Conversely, students placed into a lower level could develop an inferiority complex which could dampen their motivation to learn. As for drawbacks for instructors, Kim (2012) suggested that fine division of students into ability groups creates an unnecessarily heavy workload, as teachers must design, build, and implement syllabi across multiple classes.

Overall, while level division seems intuitively beneficial for teachers and students, the benefit for students seems to be minimal, and not universal. Hattie (2009), in one of the largest meta-analyses of studies on ability grouping, found a minimal, but statistically significant benefit in learning outcomes for students split into ability levels. Sheppard et. al (2017) found similar results in a narrower study of Japanese learners in an English for Specific Purposes (ESP) program but showed that the benefits for lower-level learners were more significant than for higher level learners.

Necessity of Level Division in Japanese Tertiary EFL Education

The benefits of level splitting are statistically questionable for learners. The maximum benefit appears to be for lower ability level students. To understand the potential benefit of level splitting at a Japanese university, we must examine the state of the English language education system in Japan to determine the level of an average student.

English language education has long been a stable feature of university curricula in Japan. In secondary schools, however, English language education has seen a revolution in recent years. In 2020, the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) issued a decree stating that, to maintain a competitive edge in Asia, the Japanese education system should prioritize the development of English language skills that equip youth to better participate in the quickly globalizing economic environment (MEXT, 2020). As a result, children from primary through secondary education have seen a large increase in the number of contact hours with English education.

A larger number of contact hours does not always result in increased English skill for all students, however. Data from MEXT show that less than 50% of high school graduates achieve the goal of CEFR A2 level English proficiency (MEXT, 2022). The same data also show of all high school graduates, roughly 45% of students elect to continue education at a four-year university. The result is a wide variety of English proficiency at university entry, with an unpredictable spread of ability within the student population. Thus, it seems reasonable to assume that there will be a large enough population of lower ability students, especially at lower-level universities, to expect reasonable benefits from level division.

Level Division Methodology

With a positive expectation for educational outcomes from level division at a lower-level university established, we examine the research to ascertain the most reasonable level division methodology for our context. The three most common methods for dividing students into levels are institutional placement tests, placement tests drawn from course textbooks, and standardized proficiency tests.

Institutional placement tests (IPTs) are the most ideal tests for level division. IPTs are tests that are designed by a specific institution, with a specific program being taught to a specific group of students at the core of the design process (Brown, 1995). Results from IPTs will group students together on a continuum of ability directly relevant to the skills to be learned in the program for which they were developed. Unfortunately, development of IPTs requires a high degree of expertise (Brown, 1989), access to a variety of quality resources (Green, 2012) and a large investment of both time and effort (Kokhan, 2012). Although IPTs are the most valid placement method of the three, the above factors render them an unrealistic solution for resource-challenged institutions.

Tests drawn from course textbooks are typically criterion referenced tests (CRTs). CRTs measure the level of a specific knowledge set a test taker possesses (Brown, 1989). For a course with few levels, in which all levels use the same textbook, or a series of textbooks with a progressive continuum of language skills, use of a textbook derived placement test would be appropriate and would most likely result in placements with high potential for positive learning outcomes. However, if a pre-made test is taken from a textbook and used out of context to place learners without consideration of how the goals of the test and of the program align, students could be placed into levels based on criteria irrelevant to course contents (Brown, 1981). As the researcher's university uses a variety of different textbooks, a single test drawn from a single textbook would not be a valid placement option.

Standardized proficiency tests are norm referenced tests (NRTs) which test overall language ability without a pre-established connection to a course, its objectives, or materials (Brown, 1995). The results of NRTs place students on a spectrum of ability as defined by the specific constructs of the test. Test suites, such as Test of English for International Communication (TOEIC) and Test of English as a Foreign Language (TOEFL), are the most used NRTs at universities in Japan, and have been argued to align with international standards for language learning, such as the Common European Framework of Reference for Languages (Tannenbaum & Wylie, 2008). Although these tests are frequently used to measure overall language skill growth, they are not entirely valid for use in level division, unless the program in which they are used has a curriculum with goals calibrated to the tests themselves, or the test and its results are carefully examined for validity within the context of the curriculum.

Each of the above level division methods has both strengths and weaknesses. In the end, however, level division and placement are rarely performed to the highest standards of validity. They are instead a procedure that operates within the financial, material, expertise, and time limitations of a specific program at a specific institution (Kokhan, 2013). This leads many universities, including the university being considered, to opt for proficiency tests. While these tests are convenient, cost effective, and have a high degree of face validity, they also demand special considerations when determining their construct validity relative to a program.

Use of Standardized Test Scores for Level Division

Standardized test scores are commonly used for level division in both ESL programs (Kokhan, 2013; Brown 1989) and EFL programs (Sheppard et. al, 2017). As discussed, the practicality of a standardized test can outweigh the potential downsides of its use. That said, Alderson, Clapham and Wall (1995) warn users that administrators must make a clear validity argument to justify the use of a test for a purpose other than that for which it was created. To make a strong argument for the all-around validity of the test before it is used, the test should be piloted, ideally with a group of students typical to those that would eventually take the test. In the absence of a pilot group, data from past tests and placements should be examined carefully to make a validity argument and identify potential improvements in the test. Messick's (1995) construct-centered framework of validity is a useful tool for establishing the validity of a given instrument, from design, to implementation, to review.

Messick's framework contains 6 different facets of construct validity, namely content, substantive, structural, external, generalizability, and consequential. The content, substantive, and structural facets are examinations of test design, item performance, test-taker engagement, and unidimensionality. Data can be gathered for arguments regarding these facets through examination of decisions reflected in test design and of how both items and test takers performed according to expectation. An argument can be made for the external facet of validity through an examination of factors outside of the test and takers themselves that may have affected the performance of either. The generalizability facet argument can only be made with several iterations of the same test administered to a variety of different test takers. This is practically possible only if the same placement test is used for multiple cohorts of students, or if multiple tests are designed using the same specifications and used congruently with two separate groups.

While all aspects of Messick's (1995) argument-based approach to validity are vital to determining the quality of an instrument for its purpose, the most important aspect for placement tests is consequential validity. Consequential validity, which refers to the validity and appropriacy of any action resulting from measurement interpretation, can only be assessed after an instrument has been used and its effect is clear to all stakeholders. In the case of a placement test, this necessitates an examination of how many students were misplaced based on results (Hughes, 2003). This can be done in three ways. First, student post-placement academic performance can be analyzed. It may be difficult to isolate placement as the most important variable in student academic performance, however, as performance is affected by a complex web of variables. Student ability, motivation, course and material difficulty, instructional methods, classroom atmosphere, and a myriad of other internal and external factors can affect student grades. Surveying students for their opinion of their own placement is a more viable option that has been used in the literature (Bradshaw, 1990). The simplest method in the literature to judge consequential validity of a placement test is to survey the teachers of the classes (Li, 2021).

Summary

The purpose of this paper was to examine the literature to determine what considerations are necessary to make judgement about the validity of an EFL placement test administered at a small, low-level liberal arts college in Japan. While there are a variety of different beliefs and arguments in the literature about the educational efficacy of level division, for both the practicality of class size and to better target the specific needs of lower-level students, level division is, at the very least, not harmful to educational outcomes. Furthermore, the current state of secondary EFL education in Japan indicates that, at the university in question, students who most benefit from level splitting will be present in large enough numbers in each cohort to justify level division. While a level test developed specifically for the institution would be the most valid, the variety of syllabi and textbook used at each level make a single institutional level test impractical. Use of a standardized test instead is acceptable, but only if the test and its results are carefully examined to make an argument for the validity of its use for level placement. In a future study, the researcher intends to build on this research by examining the results of a standardized test used for level placement over several years to determine its validity according to Messick's (1995) argument-based validity framework.

References

Alderson, J.C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Ames, C. (1992). Classrooms: Goals, structures and student motivation. *Journal of Education Psychology, 84*, 267–271.

Bradshaw, J. (1990). Test-takers' reactions to a placement test. *Language Testing 7*(1), 13–30.

Brown, J.D. (1981). Newly placed versus continuing students: Comparing proficiency. In J.C. Fisher, M.A. Clarke, & J. Schachtner (Eds.), *On TESOL '80* (pp.111–119). TESOL.

Brown, J.D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 23*(1), 65–83.

Brown, J.D. (1995). *The elements of language curriculum: A systematic approach to program development.* Heinle.

Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. *Language Testing, 14*(2), 113–138.

Green, A. (2012). Placement testing. In C. Coombe, B. O'Sullivan, P. Davidson, & S. Stoynoff (Eds.), *The Cambridge guide to language assessment* (pp.164–170). Cambridge University Press.

Hattie, J. (2009), *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* Routledge.

Hughes, A. (2003). *Testing for language teachers.* Cambridge University Press.

Kim, Y. (2012). Implementing ability grouping in EFL contexts: Perceptions of teachers and students. *Language Teaching Research, 16*, 289–315.

Kokhan, K. (2012). Investigating the possibility of using TOEFL scores for university ESL decision-making: Placement trends and effect of time lag. *Language Testing, 29*(2), 291–308.

Kokhan, K. (2013). An argument against using standardized test scores for placement of international undergraduate students in English as a Second Language (ESL) courses. *Language Testing, 30*(4), 467–489.

Krashen, S. (1985). *Input hypothesis: Issues and implications.* Longman.

Lightbown, P., Spada, N. (2013). *How languages are learned* (4th ed.). Oxford University Press.

Li, Z. (2021). Investigating the consequences of an ESL placement test. In Chapelle, C., & Voss, E. (Eds.), *Validity argument in language testing: Case studies of validation research* (pp.294–322). Cambridge University Press.

Luo, B., & Tsai, M. (2002). Understanding EFL learners in leveled and mixed classes. In *Eleventh International Symposium on English Teaching/Fourth Pan-Asian Conference, Chien Tan Overseas Youth Activity Center, Taipei.*

Messick, S. (1995) Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

MEXT. (2020). Kongou no eigo kyouiku no kaizen – jyuujitsu housaku ni tsuite houkoku ～ guroubaruka ni taiou shita eigo kyouiku kaikaku no itsutsu no teigen ～ [Report ～ The future path and improvement of English language education ～ 5 principles for English language education revolution responding to globalization]. *Ministry of Education, Culture, Sports, Science and Technology.* Retrieved from
https://www.mext.go.jp/b_menu/shingi/chousa/shotou/102/houkoku/attach/1352464.htm

MEXT. (2022). Koutou gakkou kyouiku no gennjou ni tsuite [On the current state of secondary education]. *Ministry of Education, Culture, Sports, Science and Technology.* Retrieved from
https://www.mext.go.jp/a_menu/shotou/kaikaku/20210315-mxt_kouhou02-1.pdf

Papageorgiou, S. & Cho, Y. (2014). An investigation of the use of TOEFL® Junior Standard™ scores for ESL placement decisions in secondary education. *Language Testing, 31*(2), 233–239.

Richards, J. C., & Schmidt, R. W. (2002). *Longman dictionary of language teaching and applied linguistics.* Longman.

Sheppard, C., Manalo, E., Henning, M. (2017). Is ability grouping beneficial or detrimental to Japanese ESP students' English language proficiency development? *English for Specific*

Purposes, 49. 39–48.

Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English language test scores onto the Common European Framework of Reference: An application of standard-setting methodology* (TOEFL iBT Research Report RR-08–34). Princeton, NJ: Educational Testing. Retrieved from https://www.ets.org/Media/Research/pdf/RR-08–34.pdf

大学における英語（EFL）授業レベル分けテストの妥当性について

ローランド・ロバート J.S.

抄　　録

本研究ノートは，日本のある低偏差値の大学において実施される，第二言語としての英語（EFL）授業のレベル分けテストについて，その妥当性をどう定義するかを論じたものである。大学では，英語の授業でクラス分けテストを行うことが珍しくないが，そこでテストの妥当性や基準が厳密に定義されているとは言えない。本ノートでは，日本の大学の英語教育におけるレベル分けの必要性，レベル分けを行う手段，そしてレベル分けのために標準テストを使用する際の妥当性について文献を吟味し，今後小規模大学のレベル分けテストの妥当性検証を行うにあたっての展望を示した。

キーワード：レベル分けテスト，大学の英語教育，妥当性，標準テスト