様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月 １６ 日現在

機関番号：３２５１０
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ０２９７６

研究課題名（和文）The development of test specifications based on the CEFR-J reading and listening scales

研究課題名（英文）The development of test specifications based on the CEFR-J reading and listening scales

研究代表者
ＰＡＲＫ　Ｓｉｗｏｎ（Park, Siwon）

神田外語大学・外国語学部・教授

研究者番号：００４５８６３９

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究は、日本の英語教育研究に１）実証研究、２）テスト開発、の二つの側面で貢献できる。一つ目は、第二言語習得の言語発達段階を示すにはCEFR-Jの評価尺度のより詳細な記述が必要である一方、この評価尺度に準拠した英語リーディングおよびリスニングテスト作成は可能であることが統計的に実証された点である。二つ目は、本研究の成果としてCEFR-J リーディングおよびリスニング尺度に基づいた英語テストを開発した点である。異なる英語レベルの日本人学習者を評価するために本研究で開発した3つのリーディングテストと2つのリスニングテストは、英語学習者の長期的な言語発達の評価にも活用することができる。

研究成果の学術的意義や社会的意義
国内における英語教育改革の必要性が叫ばれる昨今、英語教育のカリキュラム・デザインやテスト開発におけるCEFR-Jの有用性を探ることは意義深い。同様に、CEFR-Jリーディングおよびリスニング尺度がレベル別テストの開発のフレームワークとなり得ることが本研究によって実証された点は、学術的にも社会的にも意義がある。さらに、本研究でCEFR-Jを基に作成したテストは将来的なCEFR-J 準拠テスト開発の基礎となり、また国内の言語テスト開発に携わる研究者にとっても研究・教育上の参考となるだろう。

研究成果の概要（英文）：The current project contributes to the field of English education in Japan from two perspectives: 1) empirical research findings, and 2) test materials development. First, the empirical aspect of the project informs that while the level specifications of the CEFR-J scales require much more specifics in realizing the developmental construct, the development of level-specific EFL reading and listening tests appears feasible as the results indicted the developed items were statistically pertinent to their intended levels. Another contribution that the project makes is the tests and the item pool developed based on the CEFR-J reading and listening scales. We produced three sets of reading and two sets of listening tests of English that could be used for assessing students' English skills at differing levels. The tests may also be used for assessing the longitudinal development of students' English skills as they are level-specific with their calibrated item difficulty.

研究分野：Language Assessment, Curriculum Development

キーワード：CEFR-J　L2 Reading　L2 Listening　Language Testing

様　式　C－19、F－19－1、Z－19、CK－19（共通）

1．研究開始当初の背景

Language proficiency scales such as the Common European Framework of Reference (CEFR) and the CEFR-J can serve numerous educational purposes in second language (L2) teaching (Brown, 1996; Negishi, Takada, & Tono, 2012; North, Ortega, & Sheehan, 2010; Tono, 2013). They provide a common basis for the development of L2 programs or curricula and function as a common yardstick for the evaluation of the program or the curriculum. Similarly, the scales can be used as a reference point of learner progress at the predefined stages of long-term development. Their scaling descriptors can also be used as a benchmarked guideline for examinee performance on a standardized L2 exam; i.e., they provide an interpretive guideline for score meanings in terms of can-do's. Likewise, when adapted with more elaboration, the scales could be turned into a set of guidelines from which tests can be built to suit local testing needs (Davidson & Fulcher, 2007). This project concerned, among other uses of the CEFR-J scales, the last purpose of the CEFR-J of "serving as the framework for the standardized tests development" of English reading and listening.

The CEFR scales have been underused for test development, and the reasons are mostly due to its characteristics: illustrative and descriptive rather than normative, language and context neutral rather than specific, and comprehensive rather than complete (North, Martyniuk, & Panthier, 2010). By being such, the scales are intended to be open and flexible enough for their users to adapt and localize to suit the intended purposes better within and across different language contexts.

In addition, the scales of the CEFR are classified as user-oriented rather than constructor-oriented, making it difficult to use them as rating scales or for the development of standardized tests (Fulcher, 2010; Hulstijn, 2007; North, 2000; Weir, 2005). Such a limitation explains why there are a number of bench-marking studies to align the scores of a standardized test onto the CEFR descriptors, while it is rare to find a study that adopted or adapted the CEFR for test development. As Davidson and Fulcher (2007) argue, the CEFR scales may be used as a springboard to task and test development, but not as a set of normative guidelines that can provide a direct reference of linguistic and/or cognitive functions to be tested.

2．研究の目的

The purpose of the project was to develop level-specific tests based on test specifications for each level of the CEFR-J reading and listening descriptors. The CEFR-J reading and listening descriptors were closely evaluated to draw up a set of detailed test specifications for each level of the CEFR-J reading and listening descriptors, and level-specific tests were constructed to ensure their theoretical as well as practical efficacy as test development guidelines. This research helped not only promote the accessibility of the CEFR-J reading and listening descriptors for test development and score interpretation, but also support the validity argument for their intended use as a framework for test development.

3．研究の方法

The project was conducted in three phases: 1) analysis and specifications, 2) test construction, 3) empirical validation. The main purpose of the "analysis and specifications" phase was to analyze the CEFR-J reading and listening descriptors to identify contextual parameters, empirical measures, and cognitive processes that were essential to fill in the specifications. In the 2nd phase of "test construction", a set of reading and listening tests were developed according to the specifications developed in the 1st phase. Three reading and two listening tests were constructed with differing difficulty. In the 3rd phase of "empirical validation", the completed tests were administered to a large group of students, and the test data were analyzed using classical test as well as latent trait analyses.

4．研究成果

As we proposed earlier in our proposal, the achievements of the project can be considered from two perspectives, theoretical and practical. The theoretical achievement concerns research findings regarding the validity argument of the CEFR-J reading and listening scales as a framework for level-specific test development. On its practical side, the current project produced three reading and two listening test forms and a large pool of reading and listening test items based on the level descriptors. The theoretical as well as practical achievements of the current project therefore can be summarized as follows:

1) Test Development
   A close examination into the developmental processes of the test specifications and the actual tests helped us to examine the efficacy of the reading and listening scales for the development of test specifications and actual tests. Using the tests that we developed, we examined the

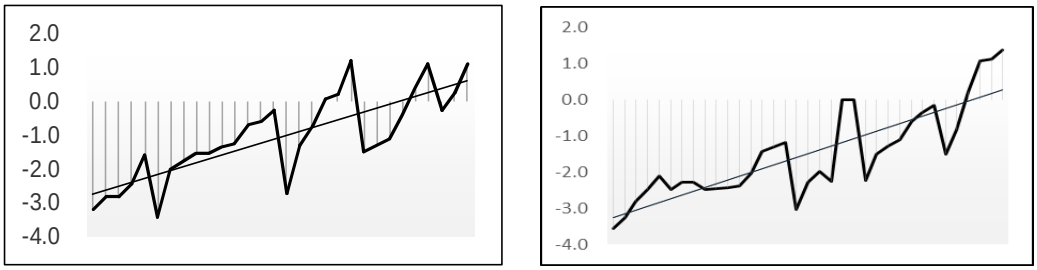
hierarchical constructs of L2 reading and listening skills depicted in the CEFR-J scales using the large test data resulted from multiple administrations of the tests. In this process, we examined the linguistic and statistical qualities of the test items and their pertinence to the scale descriptors.

2) Statistical Findings: Rank-order of the Items by Logit Difficulty

Both of the reading and listening test data were reorganized using the logit values of the items under the same sub-levels and were ordered with the items of the lower levels on the left extreme and the items of the higher levels on the right as presented in Figure 1. As it reveals, both tests include some amount of deviations in their presentation of the consecutive increment of logit difficulty. The increase of the deviation from the expected trend is clearly noticeable around the mid-levels in both tests. Nonetheless, the linear trendlines across the sub-levels in both test data exhibit the increment of logit difficulty from lower to higher levels.

Figure 1. Logit difficulty rank-order by item

[Reading] [Listening]



3) Statistical Findings: Bayesian Hypothesis Testing

Although the trendlines in Figure 1 demonstrated general progression of item difficulty in the intended and hence desirable direction, the amount of deviations some items exhibited create uncertainty as to the difficulty progression of the sublevels. That is, the mean logit values of the test items representing each sub-level (hence, sub-test) need to be evaluated for their progression of test difficulty. Therefore, Bayesian testing was performed with the reading and listening data. The test items were grouped together for each level and their mean logit values were examined using the Comparison of Means (Kuiper & Hoijtink, 2010). For the Bayesian procedures, the five sublevels were tested for their predicted implicational hypothesis; from A1.3 to B1.2 for the reading test and from A2.1 to B2.1 for the listening test. These five target sublevels were chosen to be tested as they exhibited the most amount of deviation based on the item level analyses. Hence, it was examined using the procedures if the mean difficulty at each level on each of the sampled tests increases symmetrically against the other four alternatives using Comparison of Means. The predicted hypothesis was set as $\mu 1 < \mu 2 < \mu 3 < \mu 4 < \mu 5$, which suggests that the levels present increasing difficulty from $\mu 1$ to $\mu 5$.

With the reading sublevel tests, the Bayes factor and the PMP were estimated, and the predicted hypothesis was compared against the other four alternatives using the values. Among the five hypotheses, the most supported one was the predicted hypothesis, with 46.04 of the Bayes factor and 0.39 of the PMP. Therefore, the implicational hypothesis is superior to the other hypotheses in terms of model-data fit. That is, this predicted hypothesis is empirically better supported by the data than the other hypotheses. Next, the listening tests were analyzed using the same Bayes procedures. Among the five hypotheses, the most supported one was the predicted hypothesis, with 32.74 of the Bayes factor and 0.53 of the PMP. Therefore, for both reading and listening sublevel tests, the implicational hypothesis is superior to the other hypotheses in terms of model-data fit. In other words, the ordering of mean difficulties predicted by the specifications of the CEFR-J reading and listening scales is corroborated by the empirical data obtained from the examinee participants in the current project.

4) Use of Scale Descriptors for Test Developers

A couple of issues were noted by the test developers with respects to the linguistic aspects of the scale descriptors. They found some level descriptors (e.g., A2.2 and B1.1) were not sufficient in their specificity for the text types and cognitive operations required for those levels. This finding is in line with the criticism often expressed by researchers (e.g., Weir, 2005; Fulcher, 2010). Also, the specifications frequently resort to degree words (e.g., slow, slowly, clear, or clearly) across adjacent levels. Especially in listening, the personalization of the listening stimuli (e.g., familiar to me) is common making it difficult to decide who the target learners should be. For example, B2.1 states that learners are able to read texts "within my field." However, it is not possible to know

what field the learners would be in as the tests developed based on the scale are to serve general learner population. Another issue frequently commented by the test developers were about the length of the passages or scripts, e.g., how much longer should items in the C levels be compared to the B levels. Since there is no specification regarding the length, they had to depend on their own experiences and adjust the length considering the relative difficulty of adjacent levels, making the higher level passages simply longer than the lower ones. Apparently, if a test developer continues to follow this relative approach in deciding the difficulty of test items between two adjacent levels, the scores resulted from such tests will only be interpretable in subjective terms.

5) Test Forms and Item Pool
Throughout the project, we developed three sets of reading tests and two sets of listening tests, with 28 – 33 items each. They were administered to a large group of L2 students, and using item response theory, the test data were calibrated for their parameter values that could be used for further (adaptive) test construction targeting a particular difficulty level. In addition, in the last stage of the project, we were able to develop a test item pool of 171 listening and 252 reading items representing different levels of the CEFR-J scales. These items were not administered and hence not identified for their difficulty parameter values, leaving the task as a future endeavor.

In sum, the project examined the use of the CEFR-J reading and listening scales for test development. It empirically tested the validity argument as to the CEFR-J reading and listening scales as a framework for L2 test design. A set of tests were developed through rigorous procedures to assure their quality, and the rank-order of the test items were examined using their calibrated difficulty at the item as well as test levels. These procedures informed if the scale descriptors would lead to the development of level specific tests.

In conclusion, while the level specifications of the CEFR-J scales require much more specifics in realizing the developmental construct, the development of level-specific EFL reading and listening tests appears feasible as the items rank-ordered according to their difficulty logit values demonstrated a general progression from low to high levels. Bayesian testing procedures confirmed such a progression can be valid suggesting that developing level specific tests may be possible. Finally, one needs to note that the findings in this project suggest that the feasibility of a level-specific test based on the CEFR-J scales be possible only through a rigorous coordination of the test development procedures.

５．主な発表論文等

〔雑誌論文〕（計 4 件）

(1) Megumi Sugita, Development of Level-Specific Tests Using the CEFR-J Listening Descriptors, **神田外語大学紀要**, 31, 2019, 135-153
(2) Yasuko Ito, Developing a reading test based on CEFR-J, **神田外語大学紀要**, 31, 2019, 45-57
(3) Siwon Park, Megumi Sugita, Kento Inoue, Issues in developing English reading and listening tests based on the CEFR-J scales, **言語科学研究**, 25, 2019, 57-71
(4) Siwon Park, The Effects of Test Method on L2 Reading and Listening Performance, Journal of Pan-Pacific Association of Applied Linguistics, **査読有**, 21, 2017, 45-63

〔学会発表〕（計 4 件）

(1) Siwon Park, Test development using the CEFR-J listening and reading descriptors, KATE International Conference（**国際学会**）, 2018
(2) Megumi Sugita, Yasuko Ito, Use of the CEFR-J reading scales to develop a level-specific reading test, The 16th Asia TEFL（**国際学会**）, 2018
(3) Yasuko Ito, Megumi Sugita, Development of level-specific tests using the CEFR-J listening descriptors, The 16th Asia TEFL（**国際学会**）, 2018
(4) Siwon Park, Megumi Sugita, Development of Level-Specific Tests Based on the CEFR-J Reading Descriptors, The 22nd PAAL Conference（**国際学会**）, 2017

〔図書〕（計　0 件）

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：

種類：
番号：
出願年：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1)**研究分担者**
研究分担者氏名：伊藤　泰子
ローマ字氏名：Yasuko Ito
所属研究機関名：神田外語大学
部局名：外国語学部
職名：教授
研究者番号（8 桁）：00433681

(2)**研究分担者**
研究分担者氏名：杉田　めぐみ
ローマ字氏名：Megumi Sugita
所属研究機関名：神田外語大学
部局名：外国語学部
職名：講師
研究者番号（8 桁）：70366938