SHIKEN

Volume 20 • Number 1 • June 2016

Contents

Testing and Evaluation SIG Newsletter

ISSN 1881-5537

Shiken

Volume 20 No. 1
June 2016

Fluency awareness as a way to increase speaking ability in a first-year college level English class

Dr. Jenifer Larson-Hall
larsonhall@kitakyu-u.ac.jp
Kitakyushu University

Abstract

This study reports on using fluency awareness to develop speaking ability for Japanese students over a one-year course in communicative English. Past studies on fluency and speaking rate are reviewed and classroom practices designed to promote fluency are explained. A simple test with scores that are easily calculated and understood by students for generating fluency scores is described. This test can be used by speaking teachers for a rough estimate of fluency in low-stakes classroom assessments. The statistical analyses done for this study found that students showed substantial progress in their fluency in terms of words per minute over the course of a year.

Keywords: Fluency, fluency awareness, speaking ability, speaking rate

In the decade that I have spent teaching Japanese college students English as a second language, I have consistently found that students at the college level lack fluency in speaking English. Many times when I have read these same students' writing samples I have understood that their proficiency in English is not nearly as bad as their speaking ability made it seem; students do gain an intermediate-level facility with reading and writing English from their studies in secondary school. In teaching first-year English at the university level, therefore, one of my goals has been to improve students' ability to feel confident in using the knowledge they already possess in order to speak more fluently. This paper will report on tasks that I implemented at Fukuoka Jo Gakuin University in the 2015 school year that appear to have helped students push themselves to speak more quickly and thus sound more fluent in English. The students who are described in this study are freshman majoring in English who should have an inherently high level of motivation, but I would characterize their speaking ability as beginner or low-intermediate.

In this study I will be focusing on fluency as a measurable quality of the speech sample, which Lennon (1990) called fluency in the *narrow sense* and which Segalowitz (2010) called *utterance fluency*. Being more fluent in this sense means speaking quickly but also having fewer pauses and false starts, and having pauses in appropriate places (Al-Sibai, 2004; Chambers, 1997).

How can fluency be increased?

One influential idea in the field of Second Language Acquisition (SLA) is that because of humans' cognitive abilities, there must be a trade-off between the fluency, accuracy, and complexity of utterances of non-native speakers of languages (Skehan & Foster, 2008; Wang & Skehan, 2009). For example, if the speaker is bringing attention to the task of speaking without making any grammatical mistakes (thus focusing on accuracy), their fluency may decrease as they attempt to consciously monitor their utterances. Of course, this trade-off is not always necessary; for example, for speakers for whom most grammatical structures have become proceduralized and automatic, the trade-off between accuracy and fluency would not need to occur (DeKeyser, 2007). Another way to increase the attentional resources would be to repeat a task. Research on the interaction of fluency, accuracy and complexity has found that if language learners have time to plan what they are going to say (Yuan & Ellis, 2003) or are repeating a task (Bygate, 2001) this can free up more attentional resources and fluency on the task improves.

Improved fluency, of course, is a desirable characteristic. A number of studies have found that students are perceived as more fluent speakers in general when they use a faster speech rate and have fewer pauses

(Bosker, Pinget, Quene, Sanders & de Jong, 2013; Cucchiarini, Strik, & Boves, 2002). Derwing, Rossiter, Munro and Thomson (2004) found that in their Mandarin L1 speakers' naturalistic productions that fluency was more strongly positively correlated with comprehensibility than with accentedness, even among these low-level English L2 users. In other words, judges who listened to the speech samples judged it easier to understand the speakers who they also rated as more fluent more than the speakers they rated as having better pronunciation.

One question in my mind is how fast students should be able to become in speaking. Wood (2001) reports that native speakers produce conversational English at an average speed of 270-300 syllables per minute (spm). McGuire (2009), who tested 19 English learners from a variety of first language (L1) backgrounds in a university English program, found that his control group spoke at an average of 148 spm before the experimental treatment, and his experimental group spoke at an average of 130 spm. After the treatment, which lasted for 5 weeks with three 30-minute sessions per week, the control group remained essentially unchanged (147 spm) while the experimental group increased their fluency (151 spm). Towell, Hawkins and Bazergui (1996) looked at twelve English L1 college students studying French. These students' average fluency increased over a year, from their second to third year of study of French which included a 6-month stay in a French-speaking country, from 137 spm to 157 spm. De Jong and Perfetti (2011) looked at 47 students studying English in the US at the university level over 2 weeks who performed a repetition (4/3/2) task (or not) and found that their initial speed ranged from 194-209 spm and their speed on a delayed posttest ranged from 204-232 spm.

In order to help language students become fluent in an L2, one important component is to make sure students are speaking, for skill in speaking will surely not improve unless practice in speaking is undertaken (DeKeyser, 2007). Another component is repetition. Early on in the history of SLA, Nation (1989) showed that asking students to repeat their spontaneous utterances was one way to increase fluency. He asked learners to think of ideas for a talk that they gave first for 4 minutes, then gave again but with the time reduced to 3 minutes, and finally 2 minutes (the 4/3/2 technique). In this experiment, fluency was measured by words per minute, which ranged in the eight participants from 84-196 words per minute in their final (2 minute) speech sample. Seven of the participants increased their fluency from their first version to their third.

Nation's study and others (Bygate, 2001; Lynch & Maclean, 2000) showed that asking participants to repeat their utterances allowed them to increase their fluency on that particular task. Moreover, Gatbonton and Segalowitz (2005) say that theoretically repetition of tasks which are communicative should help improve automaticity. However, there have been very few studies which have longitudinally looked at whether asking students to repeat tasks can help them to increase their fluency.

Kluge and Taylor (2000) is a report of Japanese students whose homework included taping 23-minute conversations with partners once a week over the course of an academic year (presumably 30 weeks in total). The authors report that the students are themselves surprised by their increase in fluency from the beginning to the end of the year but give no concrete numbers for their fluency.

De Jong and Perfetti (2011) is basically the only study I have been able to find which examines fluency benefits from a repeated task longitudinally, but the treatment only lasted for 2 weeks. In this study participants completed a pretest and both an immediate and delayed posttest; these were 2-min. speeches. The participants were randomly assigned to a condition, which included repeating the same speech in shorter time frames (the 4/3/2 task) or giving speeches on different topics in the progressively shorter time frames. The study did not find any increase in the articulation rate (syllables per minute) from the pretest to immediate posttest for either group (Repetition or No Repetition). For the delayed posttest there was an increase numerically but it was modest and both groups increased (Repetition group: 194 spm → 195 → 204; No Repetition group: 194 → 190 → 204). The authors speculated that the increase may have been

due to their situation since they were studying English full-time at a U.S. university and were probably increasing their English ability continuously.

The current report looks at whether my Japanese students studying English in a communicative classroom over the course of a year improved their fluency in that time period when the focus of the class was on fluency improvement and there were many repetitive tasks used in the classroom. The results of De Jong and Perfetti (2011) would seem to indicate that repetition with time reduction (the 4/3/2 task) is not necessary and that simply practicing giving speeches can lead to some increases in fluency.

However, a number of caveats are in order. First, experimental research in this area typically measures not only speech rate in syllables per minute but mean length of runs, mean length of pauses, filled pauses, dysfluencies, and so on (Wood, 2001). I did not do any of this sophisticated analysis. This classroom action report simply details the very basic measure I used for fluency, which was words per minute. I opted for a measure which the students themselves would be able to easily undertake; one of the aims of my teaching during the year was to make students aware of their own fluency and to have them focus on improving speed and not worry about their accuracy. Thus, this study is not actually one that should be compared to current experimental work on fluency in the SLA field, but instead is a report of a teaching technique that might help other teachers who want to improve their students' fluency.

The second caveat is that since there was no control group for this study I have no evidence that it was actually the repetition, or even the speaking we did in class which helped most students improve in their speaking fluency over the course of the year. It may have been that students would have improved their fluency over the year in any case, since they were studying English in college. Nevertheless, transcripts of the students' pretest and posttest speeches do show that in the main the biggest improvement was in fluency, not in accuracy (although many students also seem to exhibit gains in complexity, but that is an argument for a different paper). Another logical argument is that the time spent in class during the year (a mere 88 hours of contact) should not be enough to push students to increase their fluency without any specific emphasis on this topic.

Finally, another component to increasing fluency is motivation to do so. Common sense dictates that students who feel that increasing their fluency will produce good results will be more willing to work toward that goal. Traditionally motivation has been described as either intrinsic or extrinsic, with intrinsic motivation being that which comes from within and is spurred by a person's own desires while extrinsic motivation is external rewards for behavior that others want a person to complete (Deci & Ryan, 1985). There are newer conceptions of motivation recently (Dörnyei, 2009; Segalowitz, 2010) including one psychologist who argues that the intrinsic/extrinsic dichotomy is too limiting and that in fact motivation is much more multi-faceted (Reiss, 2012). My goal is not to explore this question here so I will just note that I did give my students several possible types of motivation.

I explained to students that native speakers would be more willing to speak with them if they increased their speech rate, emphasizing that increasing their fluency would result in a higher prestige for them. I also linked 15% of their grade to an increase in speaking rate over the semester/year. To get the full points, students were told that they needed to increase their initial number of words per minute by 10 words or more. Notice that students were not competing against a fixed fluency level that they had to reach, but rather, asked to simply improve on their own level.

Procedure

51 female students from Fukuoka Jo Gakuin participated in this action research. All of the students were first-year university students. They were required to take "First-Year English" at Fukuoka Jo Gakuin University, which consisted of four 1-hour classes every week for 11 weeks in each semester. Each

semester half of those 44 contact hours were with myself, and half were with another native speaker teacher. This resulted in a total of 88 total contact hours over the year, starting in April, 2015 and ending in December, 2015.

On the first day of the year, I asked the students to introduce themselves to me for one minute. Students did not rehearse their introduction before speaking. The introduction was done simultaneously by the whole class, who recorded themselves on their cell phones. This resulted in a very noisy minute while every student spoke into her cell phone at the same time, but no students reported that they could not hear themselves in their own recordings. Students were then asked to listen to and transcribe their recording at home, count the number of words they had said, and hand in this transcription to me at the next class period. The result of this first introduction is the baseline fluency score.

On the same day students also did two other speaking pre-tests. One was about a specific topic that we would study during the semester and about which the students would talk for one minute. For example, one topic was "List one good thing and one bad thing about the Japanese educational system." For most of these topics it would be difficult to speak fluently about the topic without some rehearsal and consideration of the topic, which was my intention. I wanted a measure of the students' fluency when talking about something very familiar to them (the self-introduction) and a floor level of fluency that they should exhibit when asked to speak in an impromptu way about an unfamiliar topic.

During the first semester we covered five topics: Educational systems, Travel, Jobs, Cultural Differences, and Fashion & Shopping. During the second semester we covered five more topics: Dating, Food & Drink, Holidays, Reading Books, and Religion. For each topic we spent 3 days covering the topic and practicing listening and speaking activities. The students were informed in their syllabus what would be their question from that topic for the speaking test so that they knew from the beginning of each unit what they would need to talk about for that topic. An example from Jobs was to imagine that they were being interviewed for a job as secretary for the Fukuoka Jo Gakuin English center and they should tell why they wanted the job and describe their fit for the job in detail.

Before speaking, students listened to utterances or speeches from native speakers which were relevant to the topic. For the second semester I had had a chance to gather unrehearsed answers from real native speakers on all of the same questions that students would talk on, but for the first semester listening activities were sometimes simply activities from the internet that pertained to the topic. Whichever type of activity that students listened to, they were asked to perform some kind of task during their first and second listening, with perhaps a third listening included if students were having a hard time answering the task questions. Finally, students were handed a transcript of what they had listened to and listened again and could follow along with the transcript. Recordings used for listening activities rarely exceeded 2 minutes.

After doing listening activities students were given time to practice speaking on the question at hand. They most often practiced this by using a speaking line where they faced a partner and spoke to their partner for 1 minute on the topic, then changed partners several more times so they could rehearse saying the same information multiple times but each time they would speak to someone who had not heard them before. This is similar to Nation's (1989) 4/3/2 technique in that the students say the same unrehearsed information multiple times, but different in that they were not asked to speak more fluently in that time.

Students would usually get a small amount of feedback on vocabulary after such activities, and I as the teacher always asked if they had any questions after the first round of speaking, trying to get them to tell me what words they had said in Japanese in their speech and needed to learn in English. I should note that this approach was not very useful as students seemed reluctant to ask for help in front of others. A technique I used once in the second semester seemed more useful: I asked the students to record

themselves doing the speaking cold at the beginning of the unit before they did any listening practice, then they transcribed that first speech. I looked at the transcriptions and after the students did the listening activities, I offered them some ways of improving on the phrases they had said. For example, when talking about what kinds of books they liked to read, I suggested that rather than say "Fantasy makes me fun" they should say "Fantasy is fun and exciting" and rather than say "This story moves my heart" they should say "It was a moving story".

At the end of the 3 class days studying a topic students would record the answers to the one or two questions in that particular topic. Usually I would have them practice their answers two or three times with a partner before recording, just to make sure they were feeling limbered up and ready to speak. Recording was done simultaneously for all students.

In both a midterm and final exam at the end of each semester students had a speaking test where I as the teacher listened to their actual speech and graded them on fluency and vocabulary. Students' speaking rate for the initial introduction fluency test was recorded on their grading sheet and they were expected to be able to speak about the topic at hand as quickly as they had been able to introduce themselves at the beginning of the semester. For vocabulary, students were expected to use some words or phrases that we had studied in the transcripts in order to sound natural in their speaking about a particular topic. Students were graded down if it seemed that they could only say whatever they would have been able to say about the topic before studying it. In most cases I didn't actually know, however, what each student was capable of for the topic before studying it; however, for the second semester topic of Reading I had their transcription of what they had said before any listening or speaking activities, and it was clear that some students were simply repeating the same things they had said before they began to study the topic. For the future in such a class I would like to work more on having students transcribe what they are naturally capable of at first and then having them notice useful phrases in listening activities and adding these to improve or change their own speaking abilities.

On the last day of class before the final listening and speaking exams, I asked students to introduce themselves once again, without any prior rehearsal. In the course of the year students had not worked specifically on improving this personal introduction. They then transcribed this and counted the number of words that they had used. This score is called here the final fluency score.

One issue that may have affected the improvement of scores from the initial to final fluency tests is that midway through the second semester I suspected that students may not have been transcribing what they said as accurately as I had hoped. In all of the spoken transcripts which I gave students I transcribed everything that the speakers said, including repetitions of the same words or false starts, but for one question in the Food & Drink section where students described how to make a particular food, I asked students to send me their speech samples and compared these to their transcripts. Most students were smoothing out their speech by not transcribing false starts or repeated words. An example of a student who had done this is shown below.

Student's transcription:

> I'd like to tell you how to make "omuraisu". You chop some vegetable and chicken. And you fry them on the frypan and season with pepper and salt. And add to rice and ketchup. Make hill and make egg crepe. (40 words)

Jenifer's transcription:

> I'd like to tell you how to make "omuraisu". You chop some vegetable and chicken. They . . uh. . . you fry them on the frypan and sea- season with pepper and salt. And add add to rice and ketchup. Mix. (laugh) Uh, make, *nandake*, make a hill? make a hill, *de*, uh, egg, egg, you make

egg, egg crepe *yatake* . . . *kazusemasu*. (64 words minus four Japanese words shown in italics = 60 words)

I showed this example to my students and asked them to transcribe all words in the future, including pause fillers such as 'uh'. I asked students to count these words because they are part of meaningful speech. I also wanted to encourage the students to use English fillers such as 'uh' and 'um', although I noted that they should only count one filler per sentence (otherwise they might easily reach an increase of 10 words by just saying 'uh' 10 times).

Clearly this change in the manner of counting words could artificially inflate the difference between the initial and final fluency tests. Out of the 51 students, only 18 wrote down any false starts, repeated words or pause fillers on their final fluency test. Among these students, the number of words added to their total score ranged from 1-7, with an average of 2.8 words. I thus conclude that although this change in approach may have resulted in a slight increase in words for the students who did it, this change was not much of a factor in their fluency increase.

Another issue in the way that I conducted this experiment is that I asked the students to transcribe their own speeches. It is possible that some students were not accurate in their transcription or even artificially inflated their number of words. Given my time with the students and the one time where I asked for sound files and checked them against what students had turned in to me I think this is unlikely, but I reiterate that this study does not claim to be an empirically unassailable report on the fluency of my students, but documents a teaching method that I used that I do think helped my students increase their fluency. I think that some students did artificially increase the number of words in their initial fluency test, in order to look better. Remember that they did this task before they knew they would have to *beat* that number. I did tell them to say as much as they could but that when they transcribed it was important not to add anything that they didn't say, but I have a few students who did not improve much over the year and from looking at their initial fluency task I suspect it was because they wrote down what they would like to have said instead of what they actually did say.

Results

Summary statistics in Table 1 for the Fluency measure show that there was a considerable increase in the students' fluency score from the beginning of the year to the end. The lowest mean score was for the topic pretest, which was in line with my assumptions. For this activity students were asked to give an impromptu speech on an unfamiliar topic, so this can be considered the students' floor level of fluency. Students were able to speak almost 15 more words per minute in their self-introduction at the beginning of year than they were for an unknown topic at the beginning of the year. For the change from the beginning to the end of the year in the self-introduction, the mean increase in words spoken per minute was about 25 words per minute, quite a large increase. The standard deviation did increase slightly from the pretest to the posttest.

Table 1
Fluency scores measured in words per minute for three tests

	Topic Pretest	Self-intro Pretest	Self-intro Posttest
Mean score	31.5	44.6	70.9
Standard deviation	15.9	19.1	23.3
N	51	51	51

Pearson's *r* correlations between fluency scores and proficiency scores found effect sizes of the correlations were large. The correlation found for the pretests of fluency scores on the Self-intro Pretest

and beginning-of-the year TOEIC Bridge proficiency scores measured by the school for all students found the effect size of the correlation was large (95% CI: .28, .70; r = .52, p < .001, N = 50, R^2 = .27). Similarly for the posttests, fluency scores on the Self-intro Posttest and end-of-the year TOEIC Bridge proficiency scores found the effect size of the correlation was large (95% CI: .31, .71; r = .54, p < .001, N = 50, R^2 = .29). This shows that there is a strong relationship between the fluency scores and proficiency scores.

Figure 1 shows a boxplot of fluency scores (measured in words per minute) from the one-minute introduction at the beginning of the first semester (Pretest) and end of the second semester of study (Posttest). Individual data points are overlaid on the boxplot.

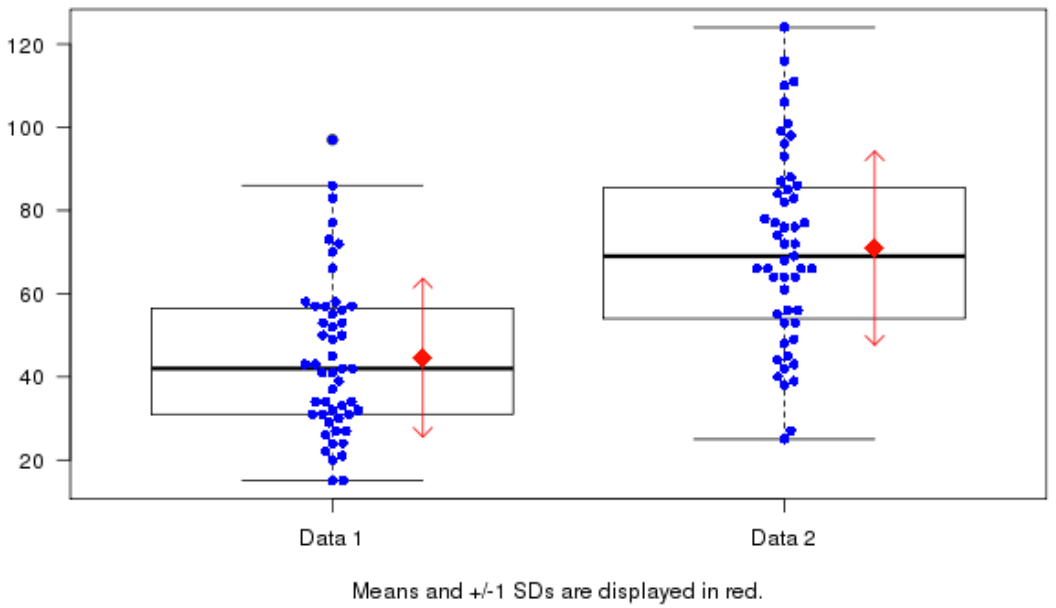

Figure 1. Scores in words per minute from pretest and posttest

Figure 2 shows the same data using parallel coordinate plots which show individual changes from pretest to posttest. This graphic shows that most students did show an upward trend, some quite steep, in their fluency scores from the beginning to the end of the semester. Only a few lines show a downward trend from the pretest to posttest. The average score is shown with the thick black line.

The boxplots show that this data is not normally distributed, as there is an outlier in the pretest data, so I performed a bootstrapped 20% means-trimmed paired-samples *t*-test on data to answer the question of whether the difference between times was statistical. This test found a 95% confidence interval for the difference between times to be 95% CI: [-30.9, -20.7]. This means that if the test were conducted multiple times, 95% of the time we could expect the true difference in the number of words produced on the pretest versus the posttest to be between 21 to 31 words, which is a large difference. Remember that in the studies examined in the literature review the largest gains were about 20 syllables per minute. Twenty-one words will be equal to at least 21 syllables at least, but usually more than that. The effect size for the difference between the pretest and posttest is d = 1.6, where I used the average of the two standard deviations as the standardizer and where the size was corrected for dependence between means.

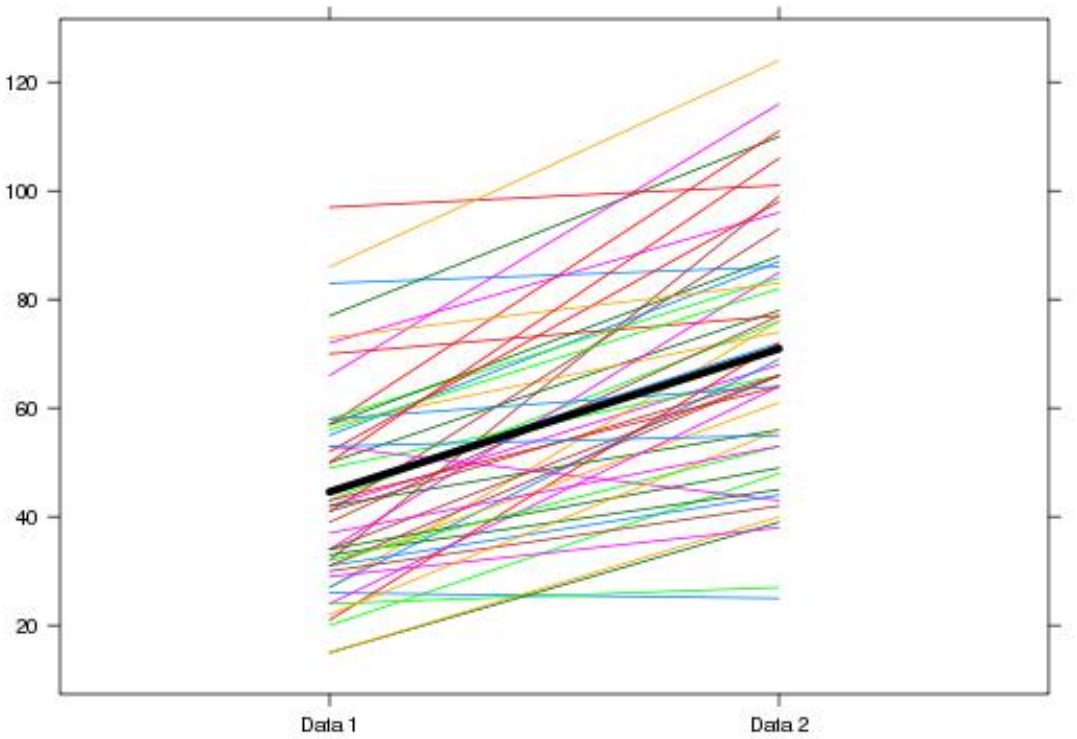

Figure 2. Parallel coordinate plots from pretest to posttest

I would like to provide a few examples of students' transcripts from the initial to the final introduction in Table 2. KI's and HK's examples are very typical of the average student who gained 20-30 words. The topics covered in the pretest are repeated, but the student is able to talk about some additional topics. Grammatical and colloquial infelicities are not corrected (KI: "my old sister") between the versions and in fact are simply repeated across the versions. MT is an additional example of a student who gained 30+ words, even though her initial speaking speed was very slow.

On the other hand, AY and YE lost words. AY's speed was quite high (well above the average) and she was able to say quite a lot in 1 minute so perhaps it is not unusual that her rate basically stayed the same over the year, and her content was basically the same too. However, looking at Figure 2 it can be seen that most of the students who started at 60 words per minute or more were able to increase their speed somewhat over the year. On the other hand, with YE I suspect that her initial introduction may have been doctored, as it seems much more polished, grammatically correct and complex than her final introduction. I did not notice many examples like YE's however, so I do not think this was a major trend in my corpus.

Table 2
Transcripts of self-introductions from the pretest and posttest

ID	Pretest	Posttest
KI	My name is K. I. I'm from Y. city in Fukuoka. I have four people and one dog in my family. My father and my mother and old sister and me. I like listen to music. My favorite singer is Miriya Kato and Taylor Swift. (45 words)	I'm K.I. I'm from Y. City. I have four people in my family. My father and my mother and old sister and one dog and me. The most favorite food is hamburg steak. It is very delicious. My hobby is shopping and watching movie. My favorite shop is ZARA and Forever21. This shop is very cheap and fashionable, so I like it. I often go to Tenjin and Hakata for shopping. (72 words)
HK	Hi, I'm H. I'm from Fukuoka. I like playing tennis very much. There are five people in my family. I have two younger sisters. I like chocolate. (27 words)	I am H. K. I am college student. There are five people in my family. I have two younger sisters, so I am the oldest children. My friends often say that I am the youngest children but it is false. I was a member of tennis club in my high school. I like playing tennis very much, so I often watch games on TV. I also like eating. My favorite food is squid and chocolate. (75 words)
MT	My name is M. T. [my name means 'green'] because I like vegetable. I like softball. Thank you. (15 words)	My name is M. T. I'm 19. My family father, mother, taller brother, sister and me. My father and mother work. My brother and sister is teacher in high school. And I'm student. I like sports. Softball and lacrosse. (39 words)
AY	Hi! I'm A.Y. I'm from A. O. A. O. is very small island that is located in south of K. There are beautiful sea and mountain. My hobby is dancing. I can dance ballet, jazz, tap, hip-hop, lock and rhythmic gymnastics. I belonged to the rhythmic gymnastics club for three years in high school. Also, I can sign language. I've been learning sign language since I was four years old. My dream is to be a flight attendant. I want to improve my English skill at this college. (88 words)	I'm A.Y. I'm from A. O. This is very small island. That located in south of K. A. is a very beautiful island. I'm really missing my home town. I like dancing. I can dance hip-hop, tap, house, lock, jazz and rhythmic gymnastics. I belonged to the rhythmic gymnastics club for three years in high school. I won the second prize in a participation. I can use sign language. I've been learning it since I was 4 years old. My part time job is Izakaya. (86 words)
YE	My name is Y. E. I'm always very fine and friendly. So I like speaking with my friends. And I like shopping and fashion. I often go to Tenjin Core or Chikushino Aeon [mall]. In the future, I hope to work at the fashion company in foreign company. And I want to be a fashion stylist. So I will study English hard. (61 words)	I'm Y. E. Um . . . I'm Fukuoka JoGakuin University student. Um . . . I come from Saitama. But I'm living Daizenji now. Um . . . I like English very much. So I study it every day. Then, I have studied it since I was elementary school student. Hm. . . In the future, I want to work in foreign countries. (53 words)

Conclusion

Judging by the summary statistics and inferential statistics, my English class progressed substantially in their fluency over the course of a year. With the average words produced per minute for the initial introduction being 44, an increase of 25 words represents an almost 60% increase in ability over the year. Also, 25 words is nearly equal to the 31 words that students could produce on an unknown topic

spontaneously, meaning that students increased their speaking ability by as much as almost a whole minute of what they were able to say in an impromptu speech on an unrehearsed topic.

This report cannot claim that any specific type of activity promoted the fluency other than practice with speaking over the year, but it is likely my emphasis on fluency, taken together with the time for practice provided in the classroom and possibly the repetition of speeches throughout the year helped lead students to increase the speed of their speech from the beginning to the end of the year.

References

Al-Sibai, D. (2004). Promoting oral fluency of second language learners literature review. Retrieved from https://iskandargoodman.files.wordpress.com/ 2013/04/promoting-oral-fluency-of-second-language-learners.pdf

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing, 30*(2), 159-175. DOI: 10.1177/0265532212455394

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 23–48). Harlow, UK: Pearson Longman.

Chambers, F. (1997). What do we mean by fluency? *System, 25*(4), 535-544.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America, 111*(6), 2862-2873. DOI: 10.1121/1.1471894

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

De Jong, N., & Perfetti, C. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning, 61*(2), 533-568. DOI: 10.1111/j.1467-9922.2010.00620.x

DeKeyser, R. (2007). Introduction: Situating the concept of practice. In R. DeKeyser (Ed.), *Practice in a second language: Perspectives from Applied Linguistics and Cognitive Psychology* (pp. 1-18). New York: Cambridge University Press.

Derwing, T., Rossiter, M., Munro, M., & Thomson, R. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54*(4), 655–679. DOI:10.1111/j.1467-9922.2004.00282.x

Dörnyei, Z. (2009). The L2 motivational self system. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 9-42). Bristol, UK: Multilingual Matters.

Gass, S. with J. Behney & L. Plonsky (2013). *Second language acquisition: An introductory course* (4th ed). New York: Routledge.

Gatbonton, E., & Segalowitz, N. (2005). Rethinking communicative language teaching: A focus on access to fluency. *The Canadian Modern Language Review, 61*(3), 325-353. DOI: 10.3138/cmlr.61.3.325

Götz, S. (2013). *Fluency in native and non-native speech*. Amsterdam: John Benjamins.

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Erlbaum.

Kluge, D., & Taylor, M. (2000, February). Boosting speaking fluency through partner taping. *The Internet TESL Journal, 6*(2). Retrieved from http://iteslj.org/Techniques/Kluge- PartnerTaping.html

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*(3), 387-417. DOI: 10.1111/j.1467-1770.1990.tb00669.x

Lynch, T., & Maclean, J. (2000). Exploring the benefits of task repetition and recycling for classroom language learning. *Language Teaching Research, 4*(3), 221-250. DOI: 10.1177/136216880000400303

McGuire, M. (2009). *Formulaic sequences in English conversation: Improving fluency in non-native speakers*. (Unpublished masters dissertation). University of North Texas, Denton, TX.

Nation, P. (1989). Improving speaking fluency. *System, 17*(3), 377-384. DOI: 10.1016/0346-251X(89)90010-9

Ortega, L. (2009). *Understanding second language acquisition*. London: Hodder Education.

Reiss, S. (2012). Intrinsic and extrinsic motivation. *Teaching of Psychology, 39*(2), 152-156. DOI: 10.1177/0098628312437704

Segalowitz, N. (2010). *Cognitive science and second language acquisition: Cognitive bases of second language fluency*. New York: Routledge.

Skehan, P., & Foster, P. (2008). Complexity, accuracy, fluency and lexis in task-based perfomance: A meta-analysis of the Ealing research. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard & I. Vedder (Eds.), *Complexity, accuracy, and fluency in second language use, learning, and teaching*. Brussels: University of Brussels Press.

Towell, R. (1987). Variability and progress in the language development of advanced learners of a foreign language. In R. Ellis (Ed.), *Second language acquisition in context* (pp. 113-127). Toronto: Prentice-Hall.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*(1), 84-119. DOI: 10.1093/applin/17.1.84

Wang, Z., & Skehan, P. (2009). Structure, lexis, and time perspective: Influences on task performance. In P. Skehan (Ed), *Processing perspectives on task performance* (pp. 155-186). Philadelphia: John Benjamins.

Wood, D. (2001). In search of fluency: What is it and how can we teach it? *The Canadian Modern Language Review, 57*(4), 573-589. DOI: 10.3138/cmlr.57.4.573

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics, 24*(1), 1–27. DOI: 10.1093/applin/24.1.1

Biographical Sketch

Jenifer Larson-Hall (Ph.D.) formerly taught at Fukuoka Jo Gakuin University where she did this research. She now teaches at Kitakyushu University. She is the author of numerous articles in research journals. She has co-authored a book with Steven Brown on Second language acquisition myths: Applying second language research to classroom teaching published by The University of Michigan Press (2012). She has recently published A guide to doing statistics in second language research using SPSS and R (2nd edition) published by Routledge (2016).

A critique of the Grade 2 EIKEN test reading section: Analysis and suggestions

Christopher Plumb[1] and Daita Watanabe[2]
cplumb@adnoc.ae, watanabe-d@shitennoji.ed.jp
1. Abu Dhabi National Oil Company Technical Institute
2. Shitennoji Junior/Senior High School

Abstract

The increased use of the EIKEN tests from a nationally used junior and high school proficiency test within Japan to a broadly used academic entrance test both in Japan and internationally means that the test requires more research and documentation than is currently available. By utilizing the EIKEN Grade 2 Reading test as an example, this paper argues that there are limitations in the test's construction, validity and documentation. The paper also briefly argues that the test suffers partially from a "washback" effect, which is related to both the test's construction and its use as an overall entrance test with only a broadly defined construct validity. The paper concludes that much more internal documentation from the test's producers is required as well as independent research from outside individuals and institutions to verify to tests overall utility.

実用英語技能検定 (英検) が日本国内の中高の英語能力テストから、日本そして国際的な入学試験として使用されることは、現在以上にテストについて研究と検証を必要とすることを意味する。例として英検 2 級のリーディングを用いることで、この論文はテスト作成、妥当性そして検証に制限があることを論じている。また簡単に、英検が washback 効果 (テストが指導に与える影響)から部分的に悪い影響を受けていることを論じている。その washback 効果はテスト作成、そして大きく定義されるテスト作成の妥当性のみ備える入学試験としての活用に関連している。この論文は英検テストの有益性を明らかにするために、第三者による個人そして外部組織による研究と同時にテスト作成者による内部検証が必要だと結論付けている。

Keywords: EIKEN, reading assessment, validity, washback effect, high-stakes testing

The Eigo Kentei or EIKEN test is Japan's most popular and widely administered test of English language proficiency (Eiken, 2016a). Supported directly by the Japanese government, it was created by the Society for Testing English Proficiency, now known as the Eiken Foundation of Japan (Eiken), in the early 1960's and has been used by all Japanese prefectures in the public education system as a benchmark. Recently some colleges and universities have used the test for adults as a placement standard at the international level (Eiken, 2016a). It is designed to be used as an alternative to the internationally popular TOEIC and TOEFL tests, but initially designed specifically for Japanese junior and high school students (Tamura, 2006). According to Eiken, the test is widely available in Japan, is offered at less than half the cost of comparable tests, and features secure administration and score reporting (Eiken, 2016b; Tamura, 2006). The test has also begun to be more readily accepted outside of Japan, as demonstrated in the case in Australia where it was initially only accepted in Queensland but later nationally (Muroko, 2014). Due to the recent expansion of the test beyond Japan's borders, Eiken has attempted to produce more English language documentation about the test's design, administration and methodology; however, little has been published so far.

The EIKEN test system has seven levels (although these are confusingly reported as five levels with two sublevels); this review will focus on the reading section of the "Grade 2" test developed for the 2015(2) year (three versions are released each year). The test is divided into four main components: listening, reading, writing, and speaking. Grammar is often regarded as a fifth component and integrated into the other sections (Nakanishi, Hayashi, Kobayashi, & Sakuma, 2010; Tamura, 2006). The Japanese Ministry of Education, Culture, Sports, Technology (MEXT) uses the Grade 2 test as a benchmark for English proficiency at the high school level upon graduation (MEXT, 2002). As using the Grade 2 test in this manner is one of the most widely used applications of the EIKEN test system, the reading component will

be critiqued from within this context. In addition, one of us is a Japanese high school teacher who has considerable experience with the EIKEN Grade 2 test. In particular, this paper will examine the reading section's format, the validity and washback effects of the test and the documentation of the Grade 2 test and EIKEN testing system in general.

Reading Section Format and Tasks

The reading test, a sample of which is available online (Eiken, 2015), is divided into four sections consisting of different reading tasks, most often requiring the test taker to engage in "cloze" exercises. Section one is comprised of 20 cloze-based short answer questions; there is a choice between four multiple choice answers, all of which belong to the same grammar constructs (adverbs for example). Section two consists of five questions where students must choose the correct word order from a spread of five words. Section three has two sub-sections, where students must select the proper answer to complete a cloze exercise within paragraphs. In one sub-section there are four choices for each question, each belonging to the same grammar construct. In the other, there are four choices, some of which belong to different constructs. Section four differs from the previous sections; it assesses comprehension of short written passages and asks contextual questions. Similar to the third section, it is divided into sub-sections. The first analyzes an e-mail message and asks three multiple-choice questions, while the remaining two contain a short article and four or five multiple-choice questions.

Aspects of the test format and tasks present themselves as targets for criticism. Though the section is designed to assess reading as a whole, almost two-thirds of the test concentrate on grammar-based questions, and do not focus on understanding meaning or what could be defined as overall comprehension. Research has shown that reading comprehension is a difficult concept to assess (Koda, 2004); however, some aspects of the test could be focused more towards assessing overall meaning by using different question types. In addition, if the test is used as a benchmark for Japanese high school students, then some of the questions could be perceived as being out of a high-school context. For instance, questions 16 and 17 relate to situations long time employees at companies would encounter (Eiken, 2015, p. 3). One could argue that question 17 is simply testing the phrasal expression "bottom line" rather than a context, but surely this can be constructed in a manner more relevant to a "real world" situation a high school student might encounter. Another case in point is the sample e-mail. The email discusses the parameters and requirements for a security company to move its office to a new location. This sample email represents an irrelevant situation for the average high-school student and should have been developed through the use of different subject matter (Eiken, 2015, p. 7). Though Eiken insists that the tests must be relevant to the test takers at each level, this is not always the case, or at the minimum is difficult to measure (Eiken, 2016b). As little has been documented about the exact nature of the design and format of the test, it is difficult to theorize about the test's design and how its questions are developed and verified.

Validity and Documentation

According to Cumming (2012), a second language examination with a high validity rating necessitates that the outcome of the test depend solely on its construct for assessment and not other points. Therefore, a tests validity requires a necessary framing of the way it conceptualizes and addresses what is language competence. This makes it difficult to discuss the validity of the EIKEN Grade 2 test, because the test has not been adequately documented. Additionally, Piggin (2011) notes that because Eiken does not adequately demonstrate what it defines as language ability and states that the test is to be used as a "broad spectrum of language ability", its construct validity is questionable. While Eiken has begun the process, it has not completed collecting the data for the construct in order to undertake a comprehensive and cohesive validity study (Eiken, 2016b). In addition, few outside researchers have studied the validity or reliability of the EIKEN test. One of the few to do so studied the EIKEN 1 test and noted that outside

research into the test has been limited and concludes that much more is required (Piggin, 2011). Sarich (2012) noted the same and went further, stating that because in practice the EIKEN test is often utilized for many different uses to measure proficiency (both inside and outside Japan), the test is not always being used for the purposes it has been designed for; as documentation related to the tests design is incomplete, this makes use of the test for different tasks more problematic. What outside research that has been done often focuses on comparisons between EIKEN Grade 1 test scores and TOEIC, TOEFL, and CEFR benchmarks (Dunlea & Matsudaira, 2009; Eiken, 2016b). Little research has been done on the other four levels, or specifically on their subsections such as reading. Additionally, and perhaps most concerning is the documentation listed by Eiken itself, which includes many references, but very few after 2008 aside from those that compare the test to the aforementioned benchmarks (Eiken, 2016b).

As part of constructing the EIKEN test's validity, Eiken published the EIKEN "Can-do List" revealing what each grade level should be able to perform, on the presumption that they are engaged in, "English in real-life situations" (Eiken, 2008, p. 4). Beginning in 2003, Eiken spent three years consulting with over 20,000 test takers to develop the Can-do list (Eiken, 2016c). In terms of reading at the Grade 2 level, the list states that students who pass the test should be able to "understand lengthy expository texts and find necessary information in texts of a practical nature" (Eiken, 2008, p. 10). Therefore, a passing score on the Grade 2 test implies the ability to read guidebooks for travelers, follow practical traffic directions, understand newspapers with Japanese explanations in footnotes, recognize the central argument of texts, comprehend sales pamphlets and easily distinguish between the topic and support sentences when reading paragraphs (Eiken, 2008). The Can-do lists represent the backbone of the EIKEN tests' benchmarking, yet very little independent research outside of the Eiken organization has been attempted (Dunlea, 2010; Nakanishi, et al., 2010). Additionally, some of the tasks referenced in the can-do lists are vague, such as stating a student proficient at the grade 2 level could "understand expository texts written for a general audience". The tasks are provided without any rationale as to what linguistic features they require or what band descriptors they may match with related standards, such as the CEFR.

In one such inquiry with students and their English instructors, Nakanishi, et al. (2010) investigated the validity of the Can-do list. They found contradictions between the list and the students' actual performance; students often could not replicate in practice the Can-do list's requirements. For instance, many learners could not accurately understand train schedules. It was theorized that this was because learners in Japan are not allocated much time for instruction focusing on authentic communicative tasks. Similarly, in anecdotal evidence provided by Piggen (2011), students commented that the test was too structured on precise details such as vocabulary, or that they could pass the test by studying test-preparation books, but the test itself was not an accurate reflection of their real abilities. Though in some cases English instructors could connect criteria taught in the classroom with the criteria of the Can-do list, the list does not discuss grammar with any criteria regarding syntax accuracy, in spite of the fact that the test, especially the reading section, contains significant sections devoted to grammar (Nakanishi, et al., 2010). Nakanishi, et al. (2010) argued the list should be adjusted to a model that considered the tasks and activities which commonly form the real curricula in Japanese classrooms. Indeed, since MEXT uses the Grade 2 EIKEN examination as a benchmark for proficiency, research should be done not only from within Eiken itself but also through various academic institutions (MEXT, 2002).

Washback

Though defined multiple ways, washback as a concept has been concisely defined as, "the impact of a test on teaching" (Wall & Alderson, 1993, p. 41). Research suggests that washback can have positive and/or negative effects on classroom instruction. For example, in Japan, teachers and students tend to focus their attention on the demands of a test, especially if that test is considered to be a high-stakes test (Buck, 1988, cited in Bailey, 1999; Piggin, 2011). This means a focus is often placed on test-taking skills, reducing

instruction on more communicative tasks. Test taking skills may include distinguishing between multiple-choice distractors or scanning for contextual clues. While it is common knowledge that Japanese English teachers often "teach to the test" with respect to the high-stakes testing - they teach specific skills that may be required for a specific test - little attention has been paid to the effects of washback concerning the EIKEN tests (Sarich, 2012).

Despite the limited research on washback effects related to EIKEN tests, there is some evidence that negative effects in the classroom have occurred as a result of its format. The reading section in particular can be singled out for such criticism. Tamura (2006) noted there was a "strategic approach to the EIKEN grammar" and that teaching to the test should be considered a reasonable strategy to pass the EIKEN test. With more than half of the reading section of the Grade 2 test being comprised of grammar-based questions, the approach seems practical. However, the effect on overall English study in the classroom could be problematic as resources and time are shifted to studying grammar instead of other areas that might improve students' overall English ability. It should be noted, however, that any effects of washback directly from EIKEN tests could be considered as much a problem of overworked instructors trying to assist their students with their future without the benefit of understanding the nature of how the tests are produced and what they are designed for. If this anomaly is a product of instruction related to taking the test, rather than simply a product of the test design itself, and is surely in need of much further research to determine its effect on both the test results and overall instruction in classrooms in Japan.

On 2 March 2009, it was reported that Ikubunkan Middle and High School in Tokyo had been illegally coaching students to pass the EIKEN tests after seeing the questions in advance. The teacher who opened the sealed envelopes in advance held special classes before the test to teach the students exactly how to pass (RakutenBlog, 2009). This represents a rare and drastic example of washback. Nonetheless, it demonstrates the lengths teachers are willing to go to prepare their students for the EIKEN tests. The high-stakes nature of the test as well as the way it is administered can be assumed as being partially responsible for the result. Eiken acknowledges that the EIKEN test may contribute to washback in the classroom, but has stated that proving the existence and effects of washback are difficult to ascertain with certainty (Eiken, 2016b). Finally, though Eiken states the test should test "English in real-life situations" (Eiken, 2008, p. 3), Piggen notes in her EIKEN Grade 1 study that the test is not reflective of real-world language and tends to focus more on the "professional world of work" (2011, p. 152) meaning that instruction to pass the test may focus in this area. This observation regarding the Grade 1 test can be applied to some of the material in the Grade 2 reading test (Eiken, 2015). This focus on professional English may have something to do with the test sometimes being used in the Japanese private sector as a means to assess English language ability for possible career advancement (Sarich, 2012). If so, it would mean that the test is not fully designed for its initial stakeholders (high school students for example) and that may mean that preparation for the test is done specifically to pass the test's components, rather than to increase overall English ability.

Limitations and Further Research

The most striking limitation of the EIKEN tests is the lack of documentation with respect to the test's validity construct. Eiken has acknowledged this limitation and has started to develop studies and commission research to document the test correctly (Eiken, 2016b). A serious concern that must be taken into consideration is that the EIKEN test has seven levels (5 main with 2 sub-levels in grades 1 and 2), that are divided into four sections. Research must be initiated to look at all aspects of each of the grade levels and each of their sections. The Can-do List has provided a reasonable start to establish the validity and reliability of the test, but as noted in the Nakanishi, et al. (2010) study it requires significant revision. It also requires research to determine what skills and linguistic requirements tasks listed in the Can-do statements are require to achieve success. Relating these to international standards such as the CEFR could

possibly make the test appear much more transparent. Additionally, especially as the EIKEN tests are being utilized more and more abroad, though Eiken has taken the initiative to begin to produce more research and documentation, due to its large scale use and often high-stakes nature, research from the institution that develops the test is insufficient. Independent research from other institutions, test developing bodies and academic specialists in the field needs to be done to corroborate findings Eiken have already disseminated and intend to in the future. Both Piggin (2011) and Sarich (2012) noted this several times in their studies. Finally, research into what possible effects washback may have in the classroom should be undertaken and above all, shared with instructors preparing students for the tests.

Conclusion

The EIKEN test has grown from a low-stakes national test to assess English proficiency into a high-stakes test, supported by the government in the public education system and used by international bodies for entrance into colleges and universities. This massive growth has not been in conjunction with research developed on the test. The developers of the test are aware of this fact and have started to implement some change and research. Once research has been completed, it can be reviewed both internally and by external experts to determine if either revision or more research is needed. As gathering validity evidence is a time consuming and ongoing process, producing new, widespread research on such a heavily utilized test is of high importance. In addition, research on other aspects of the test, such as its administration and the effects of washback on classroom instruction should be addressed as well. At present, other high-stakes tests which the EIKEN compares itself to have had significantly more research and development put into them, especially through the work of outside, independent researchers. More importantly, what limited research that has been conducted on the EIKEN system itself has been placed on evaluating the Grade 1 tests that compare similarly to the TOEFL test. The time has come for the same detailed efforts to be put into the Eigo Kentei tests at all levels.

References

Bailey, K., M. (1999). *Washback in language testing*. Princeton, NJ: Educational Testing Service.

Cumming, A. (2012). Validation of language assessments. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Malden, MA: Blackwell Publishing Ltd.

Dunlea, J. (2010). The EIKEN Can-do List: Improving feedback for an English proficiency test in Japan. In L. Taylor & C. J. Weir (Eds.), *Language testing matters: Investigating the wider social and educational impact of assessment - Proceedings of the ALTE Cambridge Conference April 2008* (pp. page numbers needed). Cambridge, UK: Cambridge University Press.

Dunlea, J., & Matsudaira, T. (2009). Investigating the relationship between the EIKEN tests and the CEFR. In N. Figueras & J. Noijons (Eds.), *Linking to the CEFR levels: Research perspectives*. Arnhem, NL: Cito, Council of Europe, and EALTA.

Eiken. (2008). The EIKEN can-do list: English translation Retrieved from http://www.eiken.or.jp/eiken/exam/cando/pdf/Eiken_CandoList_translation.pdf

Eiken. (2015). EIKEN test: Grade 2. Retrieved from https://www.eiken.or.jp/eiken/exam/grade_2/pdf/201502/2015-2-1ji-2kyu.pdf

Eiken. (2016a). History and purpose Retrieved 20 February, 2016, from http://stepeiken.org/history-and-purpose

Eiken. (2016b). Research Retrieved 20 February, 2016, from http://stepeiken.org/research

Eiken. (2016c). 英検 Can-do リスト Retrieved 20 February, 2016, from http://www.eiken.or.jp/eiken/exam/cando/

Koda, K. (2004). *Insights into second language reading: A cross-linguistic approach*. Cambridge, UK: Cambridge University Press.

MEXT. (2002). Japanese government policies in education, culture, sports, science and technology 2002 Retrieved 20 February, 2016, from http://www.mext.go.jp/b_menu/hakusho/html/hpac200201/hpac200201_2_015.html

Muroko, M. (2014, 23 October). Australia accepts Eiken certificate as proof of English ability, *The Japan Times*. Retrieved from http://www.japantimes.co.jp/news/2014/10/23/national/australia-accepts-eiken-certificate-as-proof-of-english-ability/#.VsmjYpx97mF

Nakanishi, C., Hayashi, C., Kobayashi, W., & Sakuma, S. (2010). Development of can-do statements for college students : A preliminary study. *大学英語教育学会(JACET)関東甲越地区研究年報 (JACET Kanto Region Annual Report), 6*, 14-23. Retrieved from http://ci.nii.ac.jp/els/110008664789.pdf?id=ART0009744109&type=pdf&lang=en&host=cinii&order_no=&ppv_type=0&lang_sw=&no=1456055406&cp=

Piggin, G. (2011). An Evaluative Commentary of the Grade 1 EIKEN Test. *Language Testing in Asia*, *1*(4), 1-24. doi: 10.1186/2229-0443-1-4-144

RakutenBlog. (2009). School admits leaking Eiken info since 1995 Retrieved 20 February, 2016, from http://plaza.rakuten.co.jp/odrix/diary/200903020000/

Sarich, E. (2012). Accountability and external testing agencies. *Language Testing in Asia, 2*(1), 26-44. doi: 10.1186/2229-0443-2-1-26

Tamura, A. (2006). The significance of the obtainment of the certificate in the English proficiency test (Eiken), and a proposal of an effective and efficient learning and teaching method to successfully prepare for passing: A strategy for internalizing basic Eiken grammar. *釧路工業高等専門学校紀要 (Kushiro Technical College Research Reports), 40*, 45-48. Retrieved from http://www.kushiro-ct.ac.jp/library/kiyo/kiyo40/tamura40.pdf

Wall, D., & Alderson, J. C. (1993). Examining washback: the Sri Lankan Impact Study. *Language Testing, 10*(1), 41-69. doi: 10.1177/026553229301000103

An interview with JD Brown

Jeffrey Durand
jdurand.teval@gmail.com
Rikkyo University

James Dean ("JD") Brown specializes in second language testing, curriculum design, research methods, and connected speech at the University of Hawai'i at Mānoa. He has appeared a number of times at the annual JALT conference, at JALT SIG conferences, at JALT Chapter meetings, and has been writing articles for Shiken for nearly 20 years. Thus, he has had a long connection with JALT and the Testing and Evaluation SIG. This year he is collecting his writings for Shiken and putting them into a book. The Testing and Evaluation SIG is happy to announce that SIG members will receive a free copy of this book. We also encourage everyone to attend this year's JALT conference, where JD Brown will be giving a keynote address. This interview with JD Brown takes a look at his writings for Shiken, testing and English teaching in Japan, and his upcoming JALT presentation.

You have been writing Stats Corner for almost 20 years. How have reader interests changed over the years? Has any advice changed over this time? What will be in your new book based on Statistics Corner?

Readers have had seemingly random interests over the years, so it is difficult to say if anything has changed. In language testing generally, there is now much less focus on score reliability. Score reliability depends on how a certain group of students interacts with a specific set of test items. Thus, test items that are 'reliable' for one administration may not be for another administration. Instead of looking at score reliability, there is much more concern nowadays with the validity of score use, especially since Sam Messick's seminal work on the issue. This is one of the biggest changes that has come about over recent decades.

The new book collects together most of the Shiken Stats Corner columns (41 of the original 43 appear in the book). It organizes them into two main parts: one on language testing and the other on language research. The three sections of the language testing part cover (a) strategies for test design and use, (b) item analysis techniques, and (c) reliability issues, while the three sections of the language research part discuss (d) planning research studies, (e) interpreting research, and (f) analyzing research results. The book will also include a table of contents and index, as well as one preface in English and one in Japanese, and an introductory chapter. I'm hoping that the resulting book will prove useful for any language tester or researcher. It is interesting to note that many of the articles from Statistics Corner have been widely cited by other authors inside and outside of the language testing and research fields, probably because they are accessible online and because the presentation of the concepts is straightforward and relatively easy to grasp.

Do you have any advice for those people who are interested in learning about statistics?

Learning statistics from the beginning is like learning anything new. Hence, it is important to find explanations of statistics that are easy to understand. The Statistics Corner in Shiken is one place to find such simple and straightforward presentations of statistics. More to the point, it is important to understand the basic concepts first as this helps researchers select the proper statistics for the situation in which they are working. Learning formulas is probably less important because modern statistical software packages take care of the mathematical calculations. However, understanding the statistical concepts, knowing which assumptions underlie each form of analysis, and knowing which statistics to use for which purpose

are all important. Otherwise, any statistical studies that result will probably have little meaning or even be misleading. Again, the new book should serve as one good resource for learning these sorts of things.

What trends do you see for testing or English in Japan?

I have criticized the university entrance exams in Japan for a number of years. But here I would like to be more positive because there have been many positive changes, especially with the "recommendation" system. This system has opened up admissions decisions to more kinds of information. Since multiple sources of information are much more likely than single sources of information (like single exam scores) to lead to reliable and valid decisions, the recommendation system, if handled properly, could be a very positive trend. That said, it is still up to the admissions officers involved to use the information in appropriate ways, which does not happen automatically. Nonetheless, having multiple sources of information to base decisions on is moving in the right direction.

It also seems to me that Japanese students who are returning from long stays abroad are being treated better in recent years. In many cases, these students have separate entrance exams and differing admissions requirements. I think, or at least hope, that educators and policy makers in Japan are beginning to realize that returnees have a lot to offer, even if they do not fit the traditional Japanese mold. The new admissions processes can help to insure that they are not frozen out of good Japanese universities.

University entrance exams have also improved with the inclusion of listening sections at a number of universities. Listening subtests provide additional information about the ability of the students to actually *use* the language, which is obviously quite different from the information provided by the traditional *yakudoku* tests that primarily measure the students' knowledge of grammar and vocabulary, and of course, their test-taking abilities. Importantly, the EIKEN tests are assessing all four skills (reading, listening, speaking, and writing), which provides even broader and better measurement of the students' abilities to use English. It would be nice if the university entrance examinations would also assess speaking and writing, and thereby provide more comprehensive assessment of the students' English abilities rather than just focusing on their knowledge of grammar, vocabulary and reading, and in some cases listening.

There are also some encouraging trends in Japan with regard to immersion learning of English. What I mean is that, every year, there seem to be more institutions where students are studying their content courses with English as the medium of instruction. English immersion K-12 schools for Japanese students and gaijin alike also seem to be more common. In ideal settings, students can learn in both English and Japanese. A few universities and one college that I know of also offer regular content courses in English. There are also co-teaching models in which a Japanese instructor lectures in English, while another instructor provides EFL support. I'm necessarily being vague here (even though I am thinking of specific places where I have visited and observed instruction) because I don't want to put any particular institution on the spot while these nascent trends are developing.

What will your keynote presentation be about at JALT this year? Can you give us a preview?

The presentation that I am planning for the JALT conference will examine the connection between testing and learning. More specifically, I will focus on how assessment can enhance learning. Taking a cognitive approach, the very definition of learning involves developing and increasing links in the brain through a process called myelination. My JALT presentation will consider how classroom activities and assessment can contribute to developing and strengthening those connections. Assessment should be part of the learning process, and not something added on to see what students have learned. Indeed, the very definition of assessment ought to be something like "classroom activities that provide systematic feedback." What I'm saying is that students need feedback from assessment to help focus and correct the

language practice that can then effectively strengthen the connections in their brains—the connections that are the physical manifestations of learning. I will not only consider the importance of such feedback, but also ways to improve feedback so that it better contributes to the learning processes. Or something like that. I haven't actually written my speech for next November yet, but I'm thinking it will be along the lines that I just outlined.

Thank you for the interview and all the advice you have given to Shiken readers over the years! We look forward to seeing you at JALT this year.

Thank you for visiting Hawai'i. It has been a pleasure talking with you. See you in at the JALT Conference in Nagoya.

Questions and answers about language testing statistics:

Characteristics of sound mixed methods research

James Dean Brown
brownj@hawaii.edu
University of Hawai'i at Mānoa

Question:

In Brown, 2005, you described the characteristics of sound qualitative research by discussing the importance of dependability, credibility, confirmability, and transferability. I think it would also be useful to know about the characteristics of sound quantitative and even mixed methods research. Could you address these research paradigms as well?

Answer:

You are absolutely right, Brown (2005) reviewed the characteristics of good quality qualitative research. Since then, the first part of your question was answered in Brown (2015a) which covered the characteristics of sound quantitative research. Here, I will address the second part of your question by examining the characteristics of sound mixed methods research. I will begin by reviewing my definition of what I think *research* is, as well as the key concepts in qualitative and quantitative research. Then I will turn to the issues that researchers need to address in order to produce sound mixed methods research. I will do so by explaining nine forms of legitimation and six techniques that can be applied. As I proceed through these explanations, you will see how mixed methods research includes both quantitative and qualitative methods, but also creates a research paradigm that is unique in its own right.

What is research?

In the two related columns on this topic (listed in the previous paragraph), I showed how I came to settle (in Brown, 1992 and 2004) on a single definition for research that was broad enough to include all the definitions listed in Brown (1992): research is "any systematic and principled inquiry." I also showed how quantitative and qualitative research can be systematic and principled in different, but similar ways. Generally speaking, *quantitative research* can be defended by the researcher and judged by the reader in terms of its reliability, validity, replicability, and generalizability. In contrast, *qualitative research* can be defended or judged in term of its dependability, credibility, confirmability, and transferability. Naturally, because mixed methods research systematically combines both quantitative and qualitative methods, mixed methods researchers should consider all of the issues raised in the previous two sentences for each of the research types, but should also consider the characteristics of properly combining the two types of research in such a way that it is not just a hodge-podge of quantitative and qualitative methods (sometimes referred to snidely as *multi-methods research*), but rather is a systematic and principled combination of the two research paradigms that results in a third paradigm—one that can truly be called *mixed methods research* (MMR).

How can we know if mixed methods research is systematic and principled?

We can enhance, defend, and judge the quality of MMR based on a concept called *legitimation* (Onwuegbuzie & Johnson, 2006). Brown (2014) defined legitimation as "the degree to which MMR integration of qualitative and quantitative research strengthens and provides legitimacy, fidelity, authority,

weight, soundness, credibility, trustworthiness, and even standing to the results and interpretations in MMR. Clearly, MMR investigators will want to think about legitimation in terms of how they can design their research to enhance it and thereby enhance the resulting *meta-inferences* (i.e., inferences at the MMR or integration level of study)" (p. 128).

Brown (2015b) summarized the extensive discussion originally presented by Onwuegbuzie and Johnson (2006, pp. 56-60) of the following nine subtypes of legitimation:

1. *Sample legitimation* - integrating qualitative and quantitative samples.
2. *Inside-outside legitimation* - adequately using insider and outsider perspectives.
3. *Weakness minimization legitimation* - compensating for the weaknesses in some approaches with the strengths of others.
4. *Sequential legitimation* - minimizing the effects of method sequencing.
5. *Conversion legitimation* - maximizing the effects of using both qualitative and quantitative data.
6. *Paradigmatic mixing legitimation* - combining and blending the traditions, standards, and belief systems that underlie qualitative and quantitative paradigms.
7. *Commensurability legitimation* - maximizing the benefits that accrue from switching and integrating different worldviews.
8. *Multiple validities legitimation* - maximizing the benefits that arise from legitimation of the separate qualitative and quantitative methods based on the use of quantitative, qualitative, and mixed validity types.
9. *Political legitimation* - maximizing the degree to which the consumers of the MMR value the inferences from both qualitative and quantitative methods.

Thus legitimation can be enhanced or defended in an MMR study by systematically combining samples, inside-outside perspectives, and paradigms, as well as by minimizing the effects of the weaknesses in and sequencing of different research methods, and maximizing the degree to which consumers value both qualitative and quantitative inferences, the effects of using both qualitative and quantitative data, integrating different worldviews, using separate qualitative and quantitative methods, and mixing validity types. Using some or all of these strategies to strengthen the legitimation of any particular MMR study will increase the soundness of any meta-inferences that result.

If these nine concepts seem a bit overwhelming, it may help to know that Brown (2015b, pp. 133-135) discusses six key practical techniques that mixed methods researcher can apply when trying to enhance the legitimation of their studies.

1. *Convergence* techniques examine the qualitative and quantitative data for evidence of similar conclusions.
2. *Divergence* techniques look at the data for contradictions, surprises, anomalies that could lead to new conclusions or to additional new research avenues.
3. *Elaboration* techniques examine the various data sources to see if some of them might amplify or expand on interpretations from other data sources.
4. *Clarification* techniques investigate various data sources to see if they might help understand, explain, or illuminate interpretations from other data sources.

5. *Exemplification* techniques look at various data sources for examples of inferences drawn from other data.
6. *Interaction* techniques move from qualitative to quantitative to qualitative and back to build cyclically on all five of the previous techniques.

Again, using these techniques in an MMR study will enhance its soundness, and as such, readers should look for evidence of these techniques in judging the quality of MMR studies.

Conclusion

In direct answer to your original question, the characteristics that researchers should employ to strengthen the quality of an MMR study and readers should look for in judging the quality of an MMR study are the following *forms of legitimation*: sample, inside-outside, weakness minimization, sequential, conversion, paradigmatic mixing, commensurability, multiple validities, and political forms of legitimation. To accomplish some or all of that, several *techniques* can be applied by the MMR investigator: convergence, divergence, elaboration, clarification, exemplification, and interaction techniques.

However, neither the MMR investigator nor the reader should expect all nine forms of legitimation and six techniques to be appropriate for any particular study. Instead, any decisions about the quality of MMR should be a matter of degrees. More specifically, it would help to ask how many of the forms of legitimation and techniques were applied? To what degree were they used? And, how effectively did they work together?

If you find MMR intriguing, you can explore further in Brown, 2014 and 2015b, or if you are hopelessly fascinated by MMR, some or all of the following general MMR books may prove useful: Bergman (2008); Cresswell (2003, 2009); Cresswell and Plano Clark (2007); Greene (2007); Mertens (2010); Plano Clark and Creswell (2008); Tashakkorie and Teddlie (1998, 2010); and Teddlie and Tashakkorie (2009).

References

Bergman, M. (Ed.). (2008). *Advances in mixed methods research.* Thousand Oaks, CA: Sage.

Brown, J. D. (1992). What is research? *TESOL Matters*, *2* (5), 10.

Brown, J. D. (2004). Research methods for Applied Linguistics: Scope, characteristics, and standards. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 476-500). Oxford: Blackwell.

Brown, J. D. (2005). Statistics Corner. Questions and answers about language testing statistics: Characteristics of sound qualitative research. *Shiken: JALT Testing & Evaluation SIG Newsletter, 9*(2), 31-33. Also retrieved from the World Wide Web at http://www.jalt.org/test/bro_22.htm

Brown, J. D. (2014). *Mixed methods research for TESOL*. Edinburgh: University of Edinburgh Press.

Brown, J. D. (2015a). Statistics Corner. Questions and answers about language testing statistics: Characteristics of sound quantitative research. *Shiken Research Bulletin, 19*(2), 24-28. http://teval.jalt.org/sites/teval.jalt.org/files/19-02-24_Brown.pdf

Brown, J. D. (2015b). Mixed methods research. In J. D. Brown & C. Coombe (Eds.) (2015c). *The Cambridge guide to research in language teaching and learning* (pp. 78-84). Cambridge: Cambridge.

Cresswell, J. W. (2003, 2009). *Research design: Qualitative, quantitative, and mixed methods approaches.* Thousand Oaks, CA: Sage.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Greene, J. C. (2007). *Mixed methods in social inquiry.* San Francisco: Wiley.

Mertens, D. M. (2010). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods.* Thousand Oaks, CA: Sage.

Onwuegbuzie, A. J., & Johnson, R. B. (2006). The validity issue in mixed research. *Research in the Schools, 13*(1), 48-63.

Plano Clark, V. L., & Creswell, J. W. (Eds.). (2008). *The mixed methods reader*. Thousand Oaks, CA: Sage.

Tashakkori, A., & Teddlie, C. (Eds.) (2010). *Sage handbook of mixed methods in social & behavioral research* (2nd ed.). Thousand Oaks, CA: Sage.

Tashakkorie, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches* (Applied Social Research Methods Series Number 46). Thousand Oaks, CA: Sage.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches to social and behavioral sciences.* Thousand Oaks, CA: Sage.

Where to submit questions:

Your question can remain anonymous if you so desire. Please submit questions for this column to the following e-mail or snail-mail addresses:

brownj@hawaii.edu.

JD Brown
Department of Second Language Studies University of Hawai‘i at Mānoa
1890 East-West Road
Honolulu, HI 96822
USA

Call for Papers

Shiken is seeking submissions for publication in the November 2016 issue. Submissions received by 1 September, 2016 will be considered, although earlier submission is strongly encouraged to allow time for review and revision. *Shiken* aims to publish articles concerning language assessment issues relevant to classroom practitioners and language program administrators. This includes, but is not limited to, research papers, replication studies, review articles, informed opinion pieces, technical advice articles, and qualitative descriptions of classroom testing issues. Article length should reflect the purpose of the article. Short, focused articles that are accessible to non-specialists are preferred and we reserve the right to edit submissions for relevance and length. Research papers should range from 4000 to 8000 words, but longer articles are acceptable provided they are clearly focused and relevant. Novice researchers are encouraged to submit, but should aim for short papers that address a single research question. Longer articles will generally only be accepted from established researchers with publication experience. Opinion pieces should be of 3000 words or less and focus on a single main issue. Many aspects of language testing draw justified criticism and we welcome articles critical of existing practices, but authors must provide evidence to support any empirical claims made. Isolated anecdotes or claims based on "commonsense" are not a sufficient evidential basis for publication.

Submissions should be formatted as a Microsoft Word (.doc or .docx format) using 12 point Times New Roman font, although plain text files (.txt format) without formatting are also acceptable. The page size should be set to A4, with a 2.5 cm margin. Separate sections for tables and figures should be appended to the end of the document following any appendices, using the section headings “Tables” and “Figures”. Tables and figures should be numbered and titled following the guidelines of the *Publication Manual of the American Psychological Association, Sixth Edition.* Within the body of the text, indicate approximately where each table or figure should appear by typing “Insert Table x” or “Insert Figure x” centered on a new line, with “x” replaced by the number of the table or figure.

The body text should be left justified, with single spacing for the text within a paragraph. Each paragraph should be separated by a double line space, either by specifying a double line space from the Microsoft Office paragraph formatting menu, or by manually typing two carriage returns in a plain text file. Do not manually type a carriage return at the end of each line of text within a paragraph.

Each section of the paper should have a section heading, following the guidelines of the *Publication Manual of the American Psychological Association, Sixth Edition.* Each section heading should be preceded by a double line space as for a regular paragraph, but followed by a single line space.

The reference section should begin on a new page immediately after the end of the body text (i.e. before any appendices, tables, and figures), with the heading “References”. Referencing should strictly follow the guidelines of the *Publication Manual of the American Psychological Association, Sixth Edition.*