SHIKEN

Volume 21 • Number 1 • June 2017

Contents

Testing and Evaluation SIG Newsletter

ISSN 1881-5537

Shiken

Volume 21 No. 1
June 2017

Language education pressures in Japanese high schools

Colin Mitchell
mitchell@reitaku-u.ac.jp
Reitaku University

Abstract

The Japanese education system has been in a constant state of reform, with pressure towards CLT (communicative language teaching) in language education being made as far back as the 1980s. Fast forward to 2017 and there appears to have been little change in language teaching approaches, with the traditional grammar translation remaining dominant. One reason put forward for this stifling development of the reforms is the rigid testing system, which pressures teachers to favor more traditional methods of teaching. Perhaps the most severe test is the National University Center Entrance Exam (Center Test), which is essential for many students who want to enter university. Although this Center Test has an English language section, it is not a language proficiency test. This leads to a language learning environment with a plethora of pressures from reforms, the Center Test and more traditional methods of teaching. This study used a mixed method research approach to explore the pressures of language education in Japanese high schools.

Keywords: Center Test, education pressures, MEXT reforms

As Japan increases its presence in an ever-expanding globalized environment, we must question whether the language education pressure in Japanese high schools is there to equip the students with the English skills needed to communicate internationally or rather it is simply to pass a test. The Ministry of Education, Culture, Sports, Science and Technology Japan (MEXT) certainly wishes to focus public language education on communication rather than grammar translation through its current reform plan. However, as far back as Hino (1988), the move away from grammar translation faced many challenges. Heavy criticism was given to the National University Center Entrance Exam (Center Test) which still exists in 2017 as a written multiple-choice test (Asquith, 2014; Browne & Wada, 1998; Samimy & Kobayashi, 2004). This test, although given as an excuse for grammar translation, could benefit from the MEXT communication focus reforms (Guest, 2008; Mulvey, 1999). MEXT also aims to inject culture into its communicative teaching, which grammar translation lacks. Liddicoat (2008) pointed out that this learning through culture has the potential to broaden students' cultural awareness, better preparing them for the globalized society envisioned by MEXT.

With pressure on the advancing and impending education reforms for English education in Japanese high schools, which aimed to move from a grammar and translation pedagogy "to one with a stronger emphasis on communication" (Matsuura, Chiba, & Hilderbrandt, 2001, p. 70), this study focused on the language education pressures in Japanese high schools, including pressures on ALTs (Assistant Language Teachers), JTs (Japanese Teachers) and high school students. These include the MEXT pressures to move away from grammar and focus on teaching culture and communication, as well as pressures around preparing for the Center Test.

The pressure to perform better internationally is reported by MEXT (2013) in the *Current Status and Issues of Education in Japan*, as well as the "rapidly declining birth-rate, aging population", called for Japan's education to pay "attention to international as well as domestic trends". These pressures resulted in MEXT drawing up the *English Education Reform Plan Corresponding to Globalization* (MEXT, 2014a) and moving away from more traditional grammar translation methods. This reform plan will start at the elementary level and continue up towards high schools, with the year 2020 set as the target for completion. In their reform plan, MEXT (2014a) revealed that "English education" in high school "should focus on the development of communication skills to convey ideas and feelings in English, rather than

grammar and translation". Here we are seeing a shift from the traditional grammar translation method to a more Communicative Language Teaching (CLT) focused teaching approach.

The rejection of grammar translation in favor of CLT is one that has been embraced globally. Although it has been argued that teaching grammar translation is less demanding for the teacher, it can be tedious for students. CLT may be more demanding on teachers, particularly non-native speaking teachers, but it provides students with opportunities to communicate in English (Richards & Rodgers, 2001). The MEXT reforms focused on three principles in education: "'Independence', 'Collaboration', and 'Creativity'" (MEXT, 2015, p. 6). According to the reform plan, the goals for these students was to have the "ability to fluently communicate with English speaking persons" (MEXT, 2014a, p.1). This goal was to be achieved by having "classes conducted in English with high-level linguistic activities", through "presentations, debates and negotiations" (MEXT, 2014a, p.1). With the student goal being to "fluently communicate with English speaking persons" (MEXT, 2014a, p.1) this could be described as weak CLT (Howatt, 1984). This was reflected in the Course of Study (CoS), the teaching guidelines set up by MEXT, which was revised in 2009 to be more in line with the MEXT reforms.

Introducing CLT in Japan has not been an easy task, Sakui (2004) suggested it is "a complicated issue, involving various factors such as teacher beliefs and contextual restraints" (p. 156), this was due to grammar translation being a teaching method which was regarded as being embedded into Japanese culture. Since grammar translation has not been supported by the government since the mid-1980s, Hino (1988) concluded that it is "not something that is politically imposed … but is a long established tradition which exists at a deeper level of the sociolinguist structure of Japan" (p. 45). With this in mind, the reforms by MEXT to introduce more CLT in the classroom have been described as a "re-culturing of schools, teachers and teaching conditions" (Mondejar, Valdivia, Laurier, & Mboutsiadis, 2011, p. 180).

MEXT responded to this by establishing the JET Programme which brought more foreign ALTs to Japanese public schools, sharing "their own culture to a local community in Japan, helping the country to gain personal contact with peoples of other countries" (JET, 2016). There are now more foreign ALTs than ever before, with a reported 62,000 ALTs from 65 different countries (JET, 2016). Following the research on teacher's pressures analyzed by of Browne and Wada (1998), it was found that teaching the contents of the textbook to be greatest pressure in high school. Furthermore, in Schneer's (2007) study of five of the most popular high school textbooks it was found that all of these books presented "Japanese and Western cultures as facts" which often "reinforced stereotypes and an us-and-them mentality" (p. 605). This was unfortunate since Kazufumi & Befu's (1993) empirical research showed that, "belief in *Nihonjinron* is negatively correlated with education, travel abroad and having foreign friends" (p. 100). *Nihonjinron* can be translated as "the question of the Japanese people" and is part of Japanese ideology and identity, includes elements of belief that Japan is "linguistically and culturally homogenous" (Liddicoat, 2008, p. 34). This contradicts the idea of a globalized Japan envisioned by MEXT. Therefore, the globalized reforms to expose students to more foreign people, such as foreign ALTs, and offer more cultural education, will have a negative effect on both teachers', and students' belief in *Nihonjinron*.

However, McConnell (2000) talked about Japanese students using topics of "age and marital status of one's conversational partner" being "crucial determinants of language and demeanor used during face-to-face interaction" (p. 86). These topics may not always be appropriate when communicating with non-Japanese people and therefore increase the need for cultural education. Many ALTs felt underutilized as educators and rather than reducing *Nihonjinron* through cultural education, it was reported by McConnell (2000) that ALTs sometimes felt they were in Japan "for a *gaijin* [foreigner] show – not for teaching" (p. 126).

McConnell (2000) also observed that some JTs believed that the "presence of an ALT would take away valuable time from entrance exam study as well as constantly threaten to embarrass the majority of the

Japanese teachers of language, whose spoken English skills were limited" (p. 168) The Association for Japanese Exchange & Teaching (AJET) in a survey of 936 ALT respondents in 2014 revealed that the "apparent goal of English education in Japanese schools at present is not to learn communicative English" as suggested by MEXT, "but rather to memorize textbook materials and pass exams" (AJET National Council, 2014, p. 12).

Browne & Wada (1998) believed that "the predominance of translation and teacher-fronted (i.e. non-communicative) teaching methodologies in public schools may be due to the overwhelmingly discrete point, receptive nature of the entrance exams" (p. 108). Underwood (2010) criticized language teachers in Japan regarding "their overemphasis on grammar-translation methodology and the discrete-point view of language" with a reported pressure from "the entrance examinations as highly influential on their practices" (p. 166). Asquith (2014) suggested high school English language testing still had a strong focus on "grammar and reading comprehension, with only a small section allocated to listening" (p. 49). With CLT, the teacher is teaching communicative language skills, however, the students need "to score highly on these tests" (p. 49). This can cause complications when teachers are following the MEXT reforms, but find it difficult adapt their CLT style to the learning outcome or goal. Grammar translation was still believed to be the easiest method for "preparation for tests and perceived as the method for success on exams" (Mondejar, et al., 2011, p. 181). One such exam is the Center Test which forms the first major test often used by high school students to enter university. This Center Test was deemed by Sakui (2004) as being "heavily grammar-orientated", yet acting as a "critical gatekeeping practice" to enter Japanese universities (p. 156). This resulted in Asquith (2014) commenting that English teaching in high schools was "outdated", with a "lack of variety and creativity in lessons" being a result of the Center Test (p. 49). However, competition for entering universities in Japan was fierce "and based solely on entrance examination scores" (Browne & Wada, 1998). The National Institute of Japanese Language and Linguistics lists 109 compounds regarding *taking an examination* or *juken* (Backhaus, 2014). Three such interesting terms *are juken kyoso (exam competition), juken senso (exam war) and juken jigoku (exam hell)* (Backhaus, 2014), which highlight the pressure students faced when taking the Center Test. This pressure came from parents who started to prepare their children for these exams at pre-school age. The Japan Times newspaper reported that "8 percent of 5-year-old kids in Tokyo take part in the process" (Clavel, 2014). Doing this could be a taboo topic since there was a great shame for parents whose child failed their entrance exam after taking part in this process, this was known as *zenmetsu* (crushing defeat) (Backhaus, 2014). Each year more than 500,000 candidates nationwide take the Center Test, "which will have a great impact on determining which university exams" (Guest, 2008, p. 86) they will take some two weeks later.

Samimy & Koybayashi (2004) showed justification for the high pressure of grammar teaching, as it was seen to be beneficial towards the Center Test. These results suggested there was a washback effect from the Center Test, where "CLT contradicts existing methods" (p. 204). However, Guest (2008) observed that when taking the Center Test, "the skill required to complete the task correctly might well demand an integrative approach" (p. 90). It should be noted that far more weight was placed upon the "more extended, comprehensive, integrated texts and tasks than upon discrete items" (p. 96). These extended reading passages were regarded by Mulvey (1999) as being "adult level, well-written" and "grammatically and stylistically correct" (p. 129). Guest (2008) drew the conclusion that this made the Center Test "not a grammar test" (p. 96), which was reinforced with Mulvey's (1999) analysis concluding that the Center Test offered "contextualized, task based questions" (p. 129).

The pressure from MEXT and the CoS was clearly toward the goal of communicating proficiently, and exposing students to culture, yet there was substantial pressure to teach grammar for the Center Test. Since the Center Test was so critical for students, with 82% of all Japanese high school students taking the test in 2016 (Kyodo, 2016) and a seeming mismatch of goals, this study attempts to understand the

teaching pressures of high school English teachers in Japan and how the Center Test and these CLT reforms effect such pressures on teachers in Japanese high schools.

Method

This research consisted of two types of data collection methods, quantitative and qualitative data collection. It focused on the pressures of English education in Japanese high schools, therefore it analyzed the data from English language high school teachers and students. For the quantitative data collection, two questionnaires were developed. One was designed for the students, and the other for the teachers. These can be found in Appendix A and Appendix B, respectively. The questions focused on pressures of teaching in Japanese high school, therefore variables in the participants were set. There were two main groups of participants: high school teachers in Japan; and Japanese first year university students who have passed the university entrance exam. The questions analyzed pressures on high school teachers' similar to the research by Browne and Wada (1998), using pressures suggested by MEXT (2014b) and Gorsuch (2001). This study aimed to concentrate on the reforms set by MEXT and the constitution of the Center Test. Knowing the pressures helps understand how these reforms are effecting the English language teaching in Japanese high schools.

Participants

In the teachers group, there were three types: Japanese teacher (JT), Assistant Language Teacher (ALT) and Other. Since team teaching is part of the MEXT guidelines, both JTs and ALTs were asked to participate. The participants chose whether they were a JT, ALT, or *Other*. If the participant chose *Other* they stated their teaching position. 100 teachers participated in the research, and they were all given the online teacher questionnaire. Out of the 100 teachers who took part in the questionnaire, 67 were ALTs or Other, and 33 were JTs. We can see from Table 1 that 5 of the teachers were ALTs and 33 were JTs. The 12 who selected *Other* were asked to specify their teaching position. Out of the 12 teachers who selected *Other*, eight were non-Japanese solo teachers (not ALTs). These solo teachers would still be teaching high school students for the entrance exams. However, the remaining four who selected *Other*, believed that many of their students would not be taking the entrance exams. We can assume that this reduces the amount of Centre Test teaching pressure these teachers face.

Table 1
Language Teachers Who Teach High School Students Aiming to Enter University in Japan

Answer Options	Response Count
ALT	55
JT	33
Other	12

All the participating students were Japanese first year university students studying at a national university in northern Japan. One hundred and fourteen students participated in the quantitative research, and were all given a paper based student questionnaire. They were all studying my language learning course, which was an optional course and focused on CLT. They were first year university students and had passed the Center Test. The students' responses were compared to the answers of the teachers. Therefore, the questions for the students and the teachers were largely the same in the quantitative research, the wording was changed as appropriate.

A qualitative data collection method was used with this research to analyze the responses from the quantitative questionnaire data. A small sample of teachers were selected by the researcher from the

teacher quantitative questionnaire participants. The sampled teachers were invited to volunteer some answers to a semi-structured email. These teachers were contacted via email addresses given voluntarily during the teacher questionnaire. This method was chosen since, in my experience, people can be reluctant to give out personal information such as telephone numbers. Out of the 23 high school English teachers who voluntarily gave their email addresses in the questionnaire, eight responded (four JTs and four ALTs). These teachers were questioned via email in relation to the answers given in the quantitative research, and encouraged to elaborate on those answers. The questions were made from the results of the quantitative research. The researcher probed the participants' answers to get further insight. The JTs will be identified in this study as JT1, JT2, JT3 and JT4, and ALTs will be ALT1, ALT2, ALT3 and ALT4.

Results

Quantitative results

As shown in Figure 1, the teachers and students ranked their teaching pressures, with one being the greatest pressure and four being the least pressure. For the JTs, *Communication* had the highest pressure with *Grammar* and *Preparing for the university center entrance exam* being closely matched 2nd and 3rd respectively. For the ALTs or Other, *Grammar* had the highest pressure with *Communication* and *Preparing for the university center entrance exam* being closely matched 2nd and 3rd respectively. We know from Richards & Rodgers (2001) that non-native speakers face a greater challenge to teach communication so we can expect the pressure to teach communication to be higher for JTs than ALTs in a CLT environment. Interestingly *Preparing for the university center entrance exam* ranked lower in terms of teaching pressure, but not by much. For the students, the Center Test yielded the most pressure. However, not significantly more than either communication or grammar. Perhaps most surprising was that *Teaching culture* wielded the lowest pressure for all.

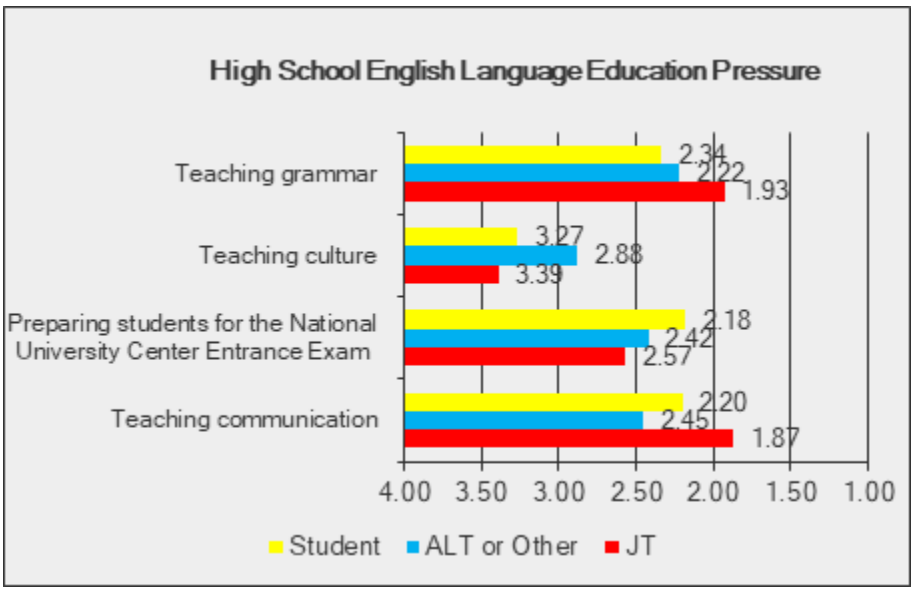

Figure 1. High school teachers and students ranking of pressure.

Qualitative results

Explanations were given by the ALTs and JTs for the lack of pressure of teaching culture in Japanese high schools. One reason was the Center Test and there being a "lack of time" (JT2) to teach the grammar and

vocabulary for the Center Test, as well as culture. JT3 believed that "high school teachers don't know how to teach language and culture", therefore teachers were giving the textbook priority, to which JT1 stated that teachers use the materials in the textbook to "let students think about culture" (JT1). However, we have also seen from Schneer (2007) that the textbooks can promote otherness. With such socio-cultural differences, conflict and misunderstandings are bound to occur. An over reliance on the textbook was further fueling the lack of culture taught in class.

Sakui (2004) acknowledged that there was a discrepancy between the definition of CLT and how to teach it. This was also attributed to a lack of time for the teacher to engage in professional development, therefore they were not getting the most up to date teaching methods proposed by MEXT and thus not feeling pressure to inject culture into their teaching. JT4 suggested many teachers still believe grammar and vocabulary memorization was the best means for high school language students.

Nihonjinron appears present as ALT3 suggested that "people like the idea" of teaching culture, however it was not seen as important since "Japan is a fairly homogenous country". This was reinforced by ALT1 who made a point that Japanese "students are not interested in other countries and cultures" (ALT1), which they believed is due to the cultural interest not being introduced. This made it difficult for ALT3 to make foreign culture relevant for students' everyday lives. ALT4 went as far to say that learning culture was "insignificant" since it was not used in testing. ALT2 also agreed with this notion.

Discussion

It appeared both JTs and ALTs in Japanese high schools supported CLT, and there was more pressure on teachers to teach communication than prepare students for the Center Test. However, grammar was much easier for the teachers to prepare. Since teachers were "expected to progress through the curriculum at a very rigid pace" (Sakui, 2004, p. 161), which included preparation for the Center Test, the pressures to teach grammar remained high. Teachers were pushed to teach from the textbook, which we have seen from Schneer (2007) can promote otherness. This can come at the expense of the ALT, who may "take away" Center Test preparation time (McConnell, 2000, p. 165).

We saw a push for teaching culture in the CoS, but this was not reflected in the classroom pressures. Young (2009) concluded that the best way to avoid excessive promotion of *Nihonjinron* was through approaching culture in an engaging and critical manner. If MEXT wants to change the approach, and move further away from the grammar method in favor of CLT, then further changes are needed. A greater focus on culture would be a major change. It is also important for teachers to realize that English is not just a "heuristic tool through which to access foreign culture" but it is also "a functional lingua franca for the exchange of ideational meaning between any members of the international community" (Seargeant, 2009, p. 60).

However, we must acknowledge that Japan is still in Kachru's (1992) expanding circle, and we cannot assume that "if something works in the inner circle […] it should work equally well in the expanding circle" (Samimy & Kobayashi, 2004, p. 249). Therefore, a simple transition from the grammar method to CLT is not going to happen. Sakui (2004) acknowledged that there is a discrepancy between the definition of CLT and how to teach it. This was also attributed to a lack of time for the teacher to engage in professional development, therefore they are not getting the most up to date teaching methods proposed by MEXT.

Conclusion

An over reliance on the textbook and underutilizing the ALT further fueled the lack of culture taught in class. While there was substantial pressure on communication, which could be seen as steps towards CLT,

there was also heavy pressure on grammar and the Center Test. Whether the washback effect of the Center Test was to blame for this or not, is more difficult to say. Certainly, it was partially to blame for its heavy grammar focus but there were many social elements involved in stifling the reforms. However, there is no reason why CLT, as defined in the CoS, cannot be used in preparation for the Center Test. Potentially it can give students a better learning experience which can prepare them for the more heavily weighted critical thinking questions.

However, the lack of pressure on culture was worrying, and suggests CLT was not being implemented in the way the MEXT reform guidelines stated in the CoS. This will do little to reduce *Nihonjinron* and will not prepare students to communicate in a globalized Japan. Giving teachers more professional development in the way of CLT training and more time to focus on their cultural teaching would be beneficial. It would be interesting to discover how much of Japanese high school language learning is devoted to studying for the Center Test, and how much those classes focus on CLT. This would reveal more about the focus of pressure of language education in Japanese high schools.

References

AJET National Council. (2014). *Assistant language teachers as solo educators.* Tokyo: AJET. Retrieved April 19, 2016, from http://ajet.net/downloads/reports/2014/ALTs_as_Solo_Educators_ENG.pdf.

Asquith, S. (2014). Integrating a functional approach with Japanese junior high school teaching practices. In P. Clements, A. Krause and H. Brown (Eds.), *JALT2014: Conversations across borders.* Ibaraki: JALT.

Backhaus, P. (2014). It's never too early to start juken. *The Japan Times.* Retrieved April 19, 2016, from http://www.japantimes.co.jp/life/2014/02/02/language/its-never-too-early-to-start-juken/

Browne, C. M., & Wada, M. (1998). Current issues in high school English teaching in Japan: An exporatory survey. *Language, Culture and Curriculum, 11*(1), 97-112. doi: 10.1080/07908319808666543

Clavel, T. (2014). Prepping for university straight from the crib. *The Japan Times.* Retrieved April 19, 2016, from http://www.japantimes.co.jp/community/2014/02/16/issues/prepping-for-university-straight-from-the-crib/#.VxWbbzB95dg

Guest, M. (2008). A comparative analysis of the Japanese university entrance Senta Shiken based on a 25-year gap. *JALT Journal, 30*(1), 85-104.

Hino, N. (1988). Yakudoku: Japan's dominant tradition in foreign language learning. *JALT Journal, 10*(1), 45-55.

Howatt, A. P. (1984). *A history of English language teaching.* Oxford: Oxford University Press.

JET. (2016). *The Japan Exchange and Teaching Programme*. Retrieved April 18, 2016, from http://jetprogramme.org/en/faq01/

Kachru, B. B. (1992). *The other tongue: English across cultures.* Illinois: University of Illinois Press.

Kazufumi, M., & Befu, H. (1993). Japanese cultural identity: An empirical investigation of nihonjinron. *Japanstudien 4*(1), 89-102. doi: 10.1080/09386491.1993.11827036.

Kyodo, J. (2016). Japan university unified entrance exams begin. *The Japan Times.* Retrieved 07 19, 2016, from http://www.japantimes.co.jp/news/2016/01/17/national/japan-university-unified-entrance-exams-begin/#.V421ODUbOls

Liddicoat, A. J. (2008). Internationalising Japan: Nihonjinron and the intercultural in Japanese language-in-education policy. *Multicultural Discourses, 2*(1), 32-46. doi: 10.2167/md043.0

Matsuura, H., Chiba, R., & Hilderbrandt, P. (2001). Beliefs about learning and teaching communicative English in Japan. *JALT Journal, 23(*1), 69-89.

McConnell, D. L. (2000). *Importing diversity: Inside Japan's JET program.* Berkeley: University of California Press.

MEXT. (2013). *Current status and issues of education in Japan*. Retrieved April 22, 2016, from http://www.mext.go.jp/en/policy/education/lawandplan/title01/detail01/sdetail01/1373809.htm

MEXT. (2014a). *English education reform plan corresponding to globalization.* Tokyo: MEXT. Retrieved March 18, 2016, from http://www.mext.go.jp/en/news/topics/detail/__icsFiles/afieldfile/2014/01/23/1343591_1.pdf

MEXT. (2014b). *Report on the future improvement and enhancement of English education (Outline): Five recommendations on the English education reform plan responding to the rapid globalization*. Retrieved March 18, 2016, from http://www.mext.go.jp/en/news/topics/detail/1372625.htm

MEXT. (2015). *Overview of the Ministry of Education, Culture, Sports, Science and Technology.* Tokyo: MEXT. 1- 36 Retrieved March 18, 2016, from http://www.mext.go.jp/en/about/pablication/__icsFiles/afieldfile/2017/02/15/1374478_001.pdf

Mondejar, M., Valdivia, L., Laurier, J., & Mboutsiadis, B. (2011). Effective implementation of foreign language education reform in Japan: What more can be done? In A. Stewart & N. Sonda (Eds.), *JALT2011: Teaching, learning, growing*. Tokyo: JALT.

Mulvey, B. (1999). A myth of influence: Japanese university entrance exams and their effect on junior and senior high school reading pedagogy. *JALT Journal, 21*(1), 125-142.

Richards, J. C., & Rodgers, T. S. (2001). *Approaches and methods in language teaching* (2nd ed.). Cambridge: Cambridge University Press.

Sakui, K. (2004). Wearing two pairs of shoes: Language teaching in Japan. *ELT Journal, 58*(2), 155-163. doi: 10.1093/elt/58.2.155

Samimy, K. K., & Kobayashi, C. (2004). Perspectives toward the development of intercultural communicative competence: Theoretical and pedagogical implications for Japanese English teachers. *JALT Journal, 26*(2), 245-261.

Schneer, D. (2007). (Inter)nationalism and English textbook endorsed by the Ministry of Education in Japan. *TESOL Quarterly, 41*(3), 600-607. doi: 10.1002/j.1545-7249.2007.tb00092.x

Seargeant, P. (2009). *The idea of English in Japan: Ideology and the evolution of a global language.* Bristol: Multilingual Matters.

Underwood, P. (2010). A comparative analysis of MEXT English reading textbooks in Japan's national center test. *RELC, 41*(2), 165-182. doi:10.1177/0033688210373128

Young, T. J., Sachdev, I., & Seedhouse, P. (2009). Teaching and learning culture on English language programmes: A critical review of the recent empirical literature. *Innovation in Language Learning and Teaching, 3*(2), 149-169. doi: 10.1080/17501220802283178

Appendix A

Student questionnaire

1. Please rank the extent of pressure 1-4 to learn the following: (1 = Greatest Pressure)

下記のことを学ぶにあたって、比重の大きい順に順位付けし○をつけてください。

（１＝一番比重が大きい）

Learning Communication コミュニケーションを学ぶこと	１・２・３・４
Preparing for the entrance center exam 大学入試センター試験対策	１・２・３・４
Learning culture 文化を学ぶこと	１・２・３・４
Learning grammar 文法を学ぶこと	１・２・３・４

Appendix B

Teacher questionnaire

1. Do you teach high school students who are aiming to get into university?

Yes – ALT	
Yes - JT	
Yes – Other (Please Specify)	

2. Please rank the extent of pressure to teach the following: (1 = Greatest Pressure)

下記について、授業で比重を置いている順に番号をつけてください

（１=最も比重を置いている)

Teaching Communication コミュニケーション英語	１・２・３・４
Preparing for the entrance center exam 大学入試センター試験対策（英語）	１・２・３・４
Teaching culture 英語圏の文化について教えること	１・２・３・４
Teaching grammar 文法を教えること	１・２・３・４

Appendix C

JT qualitative responses

Teaching culture posed the least amount of pressure for Japanese teachers and ALTs. It also ranked as having a low amount of focus in high school English class. This seems to go against the current MEXT course of study, which suggests high school English should help deepen students understanding of 'language and culture'. Why is there such a low amount of pressure and focus on teaching language culture in high school?

JT1

If you check Japanese textbook, you know this is why. Most textbooks show us about many topics related to environment, history, language, society etc. Thus, students learn not only culture but knowledge. If students learn culture, what do you teach? If so, is there enough time to teach them?

I believe that content of textbooks is not tell about culture but teachers use that materials let students think about culture.

JT2

The biggest problem is a lack of time. Even though we want to teach to help students to understand language and culture, we actually only have time to go through the textbook. I explain the culture when there is a topic relating to culture. However, it sums up to a small amount of time. If there is much time, teachers will be able to have time talking about cultural thing.

JT3

I think most high school teachers don't know how to teach language and culture. When they were students, their teachers made them memorize English grammar, words, etc not other culture, because it has nothing to do with the centre test. Also, I think most of them haven't studied abroad for a long time. After being a teacher, they don't have enough time to study for themselves because they have many things to do in a day such as preparing for classes, grading, club activities and so on. For these reasons, they don't know how to teach culture to students

JT4

I think one reason is that there remains awareness that to study English is to memorize grammar and words. In particularly, the high school only for entering universities have to focus on them. So I think the tendency is partly related to the centre test or the regular written test in the school.

Appendix D

ALT qualitative responses

Teaching culture posed the least amount of pressure for Japanese teachers and ALTs. It also ranked as having a low amount of focus in high school English class. This seems to go against the current MEXT course of study, which suggests high school English should help deepen students understanding of 'language and culture'. Why is there such a low amount of pressure and focus on teaching language culture in high school?

ALT1

Japan and Japanese people are very proud of their country. Many agree that Japan is possibly the best country in the world. Many students are not interested in other cultures or other countries. Even students that really enjoy English may have no ambition to travel outside of Japan.

“If you could go anywhere, where would you go?” -- “Tokyo Disneyland! USJ! Okinawa!”

It is possible that with interest in visiting other countries being so low, people may not see a need to introduce them. In contrast, if other cultures and countries were introduced earlier on, it could spark interest.

ALT2

Because Japanese teachers and some ALTs lack the forethought to see the importance of knowledge. Knowing why some speak, dress, live, etc the way they do will help students realize that there are other cultures out there that live differently and possibly better then Japanese. This is hard for Japanese to consider and are ashamed when it is true. After this, most teachers feel if the content will not be on the test, what is the point of learning it? Therefore, it is given little consideration and therefore lack the pressure to be taught.

ALT3

I think people like the idea, but there's simply not time for it from the teachers' perspectives. Also I don't think a lot of people see it as very important or practically useful. Because Japan is a fairly homogenous country, I don't think many people see cultural understanding as very relevant to their everyday lives.

ALT4

The teachers are more focused on the students passing exams and getting good grades. In order to do that, grammar and vocabulary must be studied and culture learning is seen as being insignificant since it generally isn't used in grading or exams.

A general overview of English tests administered in Japan[1]

Michihiro Hirai
mjhirai@beige.ocn.ne.jp
Kanagawa University

Abstract

As a means of communication, language reflects all aspects of human thoughts and activities; hence there are countless approaches to, and forms of, language tests as a means of assessing communication skills, depending on the purpose, domain, and other factors. I discussed various features and characteristics of the major English tests administered in Japan by genre such as general, academic, and purpose-specific. Just as it is important to understand what knowledge and skills each test is designed to evaluate and how, so is it also important first to realize the meaning of the test to the organization using it – be it a school or an enterprise – and then to ensure that the test serves that organization's objectives and priorities.

Keywords: testing and evaluation, correlation

Having spent more than 30 years as an engineer and a subsequent 10 years or so as a college English teacher, I have developed a grass-roots framework for discussing language education and testing, which combines the users', teachers' and learners' viewpoints. From this trilateral perspective, I have been observing the English test landscape in Japan by personally taking more than 50 English qualification tests open to the public during the past 40 or so years. In this three-part presentation, I first reviewed various parameters of language testing, then gave a comparative overview of the major English tests, and finally discussed some common misperceptions and misuses of these tests, primarily from the standpoint of industry. Considering the readership of this publication, here I am going to focus on the second and third parts, skipping the first part that discussed what is actually tested and how.

Major English tests

More than 50 different English tests are currently administered in Japan (ELT Services Japan, 2017), depending on how they are counted. ELT Services Japan's website titled 英語教育ニュース (soon to be closed down) lists 63 English tests as of the beginning of 2017, including composite tests (i.e., a combination of English and non-language knowledge/skills such as typing, accounting, and export/import trading) and tests for children, while missing many translation tests and brand-new tests such as the TEAP, not to mention minor ones such as the Pitman™ ESOL. For the purposes of this paper, I classify them into three broad categories: general, academic, and purpose-specific.

General English tests are those designed to cover a general range of use of the language and include Eiken (英検), Kokuren Eiken (国連英検), the Cambridge Main Suite (KET, PET, FCE, CAE, and CPE), IELTS™ (General Training), the Pitman ESOL, and the G-TELP™. Most of them, except the G-TELP, which does not have speaking or writing components, test all the four skills as well as grammar knowledge and test speaking skills in a face-to-face (real-person) interview mode. Most of them employ different test forms for different levels (i.e., each level or grade has its own set of questions, problems, and/or tasks), while IELTS uses a single test form to cover the entire range of levels called bands (i.e., one set of questions, problems, and tasks covers all the levels or grades) (Note: In IELTS's speaking section, the interviewer tailors test questions to the individual candidate).

[1] This report is a summary of an oral presentation at the JALT Hokkaido Conference on October 2, 2016.

Academic English tests are those designed to evaluate how capable the candidate (test-taker) is of keeping up with study at universities in English-speaking countries, while the exact acceptance criteria are usually left to individual universities. These tests include the TOEFL® iBT®, IELTS (Academic), and the GRE® General, the TEAP (Test of English for Academic Purposes), and its computer version the TEAP CBT. Historically, the TOEFL, developed by Educational Testing Service (ETS®), was practically the only major English test specifically designed for academic purposes, particularly for admission to American graduate schools. In the late 1990s, IELTS was split into General Training and Academic, and then the GRE General was introduced. Each of these tests employs a single form to cover the entire range of competence.

Recently, in response to the Japanese government's initiative to internationalize the Japanese educational system, Professor Kensaku Yoshida of Sophia University has led an ad-hoc team in cooperation with the Eiken Foundation to develop the TEAP. It was officially introduced in 2014 with only listening and reading components but now covers all the four skills and is expected to be widely used in Japan as an alternative to the English component of the National Center Test for University Admissions (センター試験).

Purpose-specific tests are very diverse in nature but can be divided into four subcategories: business or workplace English, English for tourism, technical English, and translation tests. Business or workplace English tests are those designed to evaluate the candidate's linguistic competence in business situations or at workplace, with varying degrees of business flavor. These include the TOEIC, BULATS (Business Language Testing Service) English, GTEC, BETA, Nissho Business Eiken, and TOBiS.

The TOEIC was developed by ETS in 1979 at the request of some representatives of Japanese companies, who were not quite satisfied with the Eiken test, which focused on school English (McCrostie, 2010). Their original intent was apparently to develop a more practical English test reflecting the kind of English actually used in general society. Since a number of major Japanese enterprises jumped on this bandwagon, the TOEIC has become widespread during the past three and a half decades. The popularity of the TOEIC, however, is essentially limited to Japan, South Korea, and Taiwan. The British presence is stronger in the rest of Asia as well as in Europe. In the meantime, ETS, which is a respectable organization with solid and professional research staff, has taken criticisms (Hirai 2002; Chapman, 2003) from outside seriously and has enhanced the TOEIC several times since 2006 by introducing speaking and writing components and adding a business flavor.

BULATS is available in four European languages: English, French, German, and Spanish. BULATS English was developed by Cambridge ESOL in the late 1990s, specifically as a business English test and is now one of the mainstream business English tests in Europe; however, it is not as well-known as the TOEIC in Japan. From my own experience of taking many English tests, I would rate it as the most business-oriented of all the tests.

Tests of English for tourism are, as the name suggests, mainly interpretation tests focusing on how to interact with foreign visitors and include the National Licensed Guide, the Travel English Test, and the Tourism English Proficiency Tests. In these tests, while the language content is not very challenging, a broad, often meticulous knowledge of Japanese culture and history is required, as well as some familiarity with the tourism industry.

Technical English tests are generally intended to evaluate the candidate's competence in technical communication in English and include the Waseda-Michigan Technical English Proficiency (TEP) and the Kougyou Eiken. These two tests employ distinct test forms for different levels. They both require a basic familiarity with general science and engineering, as well as a knowledge of specialized fields. In this regard, they seem similar to technical translation tests but there is a fundamental difference: The

writing components of these tests, especially the TEP (at its highest level), place great emphasis on report structure and rhetoric and are therefore very challenging in their own ways.

The term "technical English" is a major misnomer. Traditionally, technical English used generally to deal with the kind of English used in the scientific and engineering community. For this reason, it has long been made light of in English education, and its tests have been unduly unpopular. In a sense, its name has served as a self-inflicted fetter. However, the principles, particularly the writing principles, taught in technical English have recently gained international recognition as methodologies and guidelines universally applicable to all sorts of professional English. As a result, the more appropriate term 'professional English' is gaining currency and drawing more general attention.

Finally, there are a variety of translation tests including the Software Translator Test, the Intellectual Property Test, the Translator Qualifying Examination (TQE), the Hon'yaku Kentei offered by the Japan Translation Federation (JTF), the Certified Professional Translator Test offered by the Japan Translation Association (JTA), and certification tests (in various language pairs) offered by the American Translators Association (ATA). In addition, if one stretches the definition of the term "translation," the Test of Business Interpreting Skills (TOBiS) can also be grouped in this category.

Most translation tests take a single form consisting of several tasks, and the candidates are given grades. In general, these tests have very rigorous scoring criteria and focus on very discrete points (in other words, are very nitpicky). As a result, achieving the highest grade is a big challenge in any translation test, even for holders of the first grade in Eiken or the TEP, or those with a score of 900 or more in the TOEIC. It should also be noted that translation tests differ from technical writing tests in that the former demand strict adherence to the original text and proper choice of words.

Evaluating the English tests

In discussing linguistic competency from the viewpoints of knowledge and skills, I have been using a five-axis radar chart, shown in Figure 1, in which the axes *GV*, *R*, *L*, *S*, *and W* represent grammar and vocabulary, reading, listening, speaking, and writing, respectively. Using this radar chart, I would like to illustrate how the characteristics of a test affect the resulting perception of the candidate's ability.

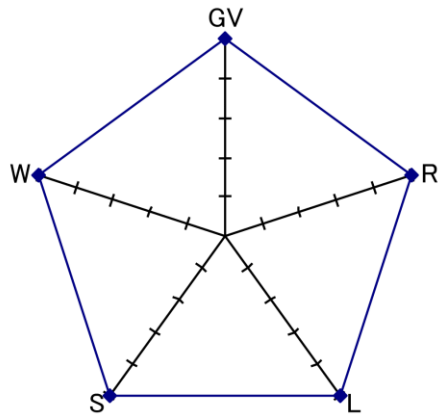

Figure 1. Radar chart of language skills.

Suppose the candidate has the ability profile depicted in Figure 2, which shows a relative strength in writing. If the test he/she takes has the evaluation profile (i.e., how thoroughly it evaluates each skill)

shown in Figure 3, which tests *GV*, *R*, and *L,* but not *S* or *W*, as is the case with the conventional TOEIC, then his/her results will have the profile shown in Figure 4, since each axis will have a value equivalent to the product of its corresponding values in Figure 2 and in Figure 3. This means that his/her speaking and writing skills may not be reflected in the test results. In other words, the test itself serves as a filter to the real ability. Here lies one of the fundamental problems inherent in language tests. The great majority of people – most critically the stakeholders – tend to use the test results as the sole source of information for assessing the candidate's linguistic competence, yet the perceived ability does not reflect his/her real ability. If the test does not adequately test active (productive) skills such as speaking and writing, or even worse, does not test them at all, then those with good active skills but relatively weak passive (receptive) skills are significantly handicapped in the test and may eventually fail it. Conversely, those who excel in grammar and passive skills but have poor active skills have a better chance of scoring high in such tests and hence are more likely to be accepted into a renowned university or assigned to a well-rewarded position in a company. This misinterpretation of test results can have significant consequences especially in the case of high-stakes tests.

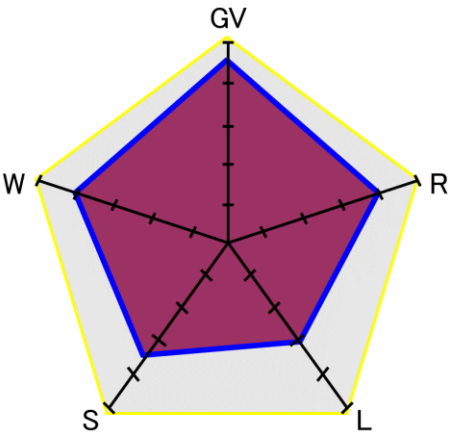

Figure 2. Radar chart of test candidate with strength in writing.

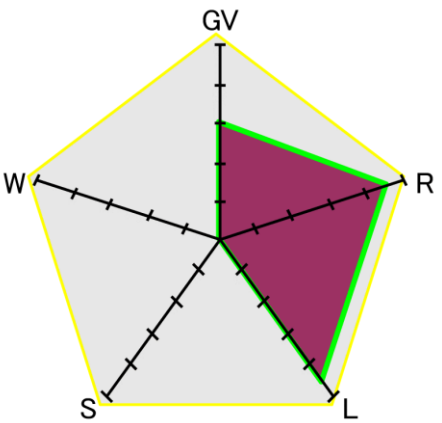

Figure 3. Evaluation profile of test lacking assessment of writing or speaking.

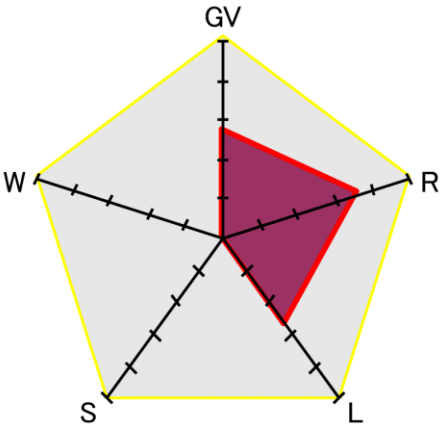

Figure 4. Radar chart of test candidate assessed by test lacking assessment of writing or speaking.

Whereas the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) (Council of Europe, 2001) provides a set of guidelines for evaluating language skills, it is no easy task to objectively determine how thoroughly each test evaluates different skills and also how difficult it is for most learners of English, primarily because most test organizations do not publish the actual test questions/problems, and also partly because test forms change from administration to administration. Based on a collection of personal memos on which I have jotted down my observations and impressions of the tests I have taken, I demonstrated in my presentation the profiles of several major tests as I perceive them. Table 1 summarizes my subjective comparison of difficulty level for some well-known English tests. In my view, among the dozens of general and academic English tests, the Cambridge Main Suite (especially the highest-grade CPE) is the most thorough and rigorous, whereas in the business and workplace domain, the BULATS English best reflects the actual needs of the business community.

Table 1
Level Comparison

CEFR	Eiken	Cambridge Main Suite*	IELTS Band	BULATS Std	BULATS S, W	TOEIC R, L	TOEIC S	TOEIC W	TOEFL PBT	TOEFL iBT
C2		CPE (A)	8.5 - 9.0	90 - 100	5					116-120
		CPE (B)	7.5 - 8.0			965-990	8	9	652-677	
	1	CPE (C)				940-960	7	8	640-651	111-115
C1		CAE	6.5 - 7.0	75 - 89	4	875-935	6	7	600-639	100-110
B2	pre- 1	FCE	5.0 - 6.0	60 - 74	3	700-870	5-6	6-7	540-599	76-99
	2					500-695	4	5-6	470-539	52-75
B1		PET	3.5 - 4.5	40 - 59	2	400-495	3-4	4-5	400-469	32-51
	pre-2									
A2	3	KET	2.0 - 3.0	20 - 39	1					
A1	4 - 5		0.0 - 1.5	0 - 19						

* For each test, the results are given in 4 grades. There may be overlaps between adjacent tests.

Note. This table summarizes the author's subjective evaluation of English tests.

Testing and learning: What they mean to business and industry

Let me stretch this pentagonal model further to illustrate how the choice of test affects the learner's progress. Figure 5 illustrates two cases, one with a passive-skills-only test (Test A) and the other with a four-skill test (Test B). It is in our nature that given any test, we tend to study only the subjects covered that it covers, especially when we are already overloaded with daily office work. Thus, with Test A, the learners are likely to stop studying or practicing active skills and, after a few years, to end up having improved only their passive skills, while letting their active skills deteriorate. On the other hand, with Test B, they will keep studying and practicing all the four skills and, after the same period of time, will have improved them all in a well-balanced manner. Accordingly, the choice of test critically molds the learners' skill profiles in the long run. This is what I call the "clothes make the man syndrome." (This phenomenon is also well known as the "washback" effects in the language testing community (e.g., Cheng and Watanabe, 2004)).

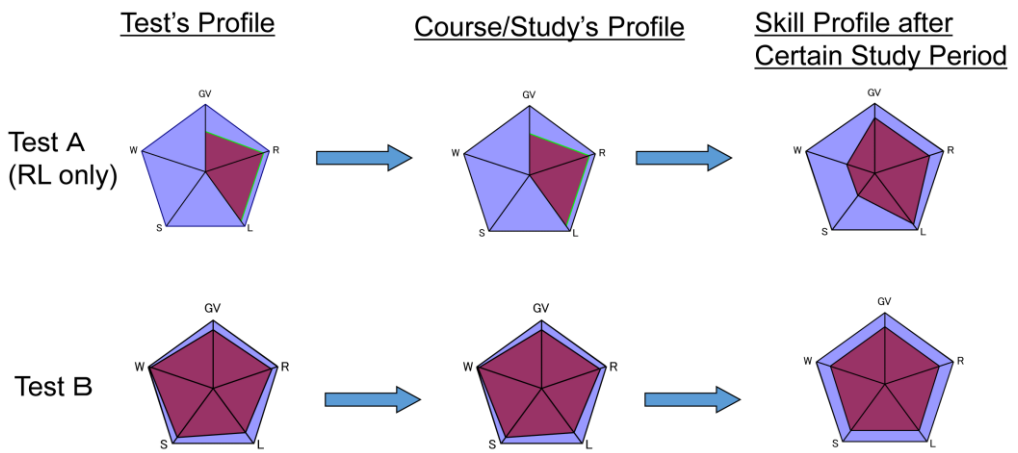

Figure 5. Washback effect of passive skills and four skills tests.

It is alarming that the prolonged use of, or dependence on, a passive-skills-only test such as the conventional TOEIC might eventually drive learners away from active skills. This is particularly true in Japanese companies, which are in great need of employees who can competently communicate with their international counterparts on the business front. A number of studies have revealed that active skills such as those required for presentations, negotiations, meetings, email writing, and report writing are high on the wanted list.

With that in mind, I have been proposing a T-square approach to corporate language education as shown in Figure 6. The idea is to build an elite pool of employees equipped with appropriate active skills (on the vertical axis) while gradually raising the average level of home-front staff on a long-term basis (on the horizontal axis). It is important to note that in language, quantity cannot substitute for quality, in other words, no matter how many 2nd grade speakers a company may have, it cannot beat a team comprising one top-grade person. In this respect, the human resources department should carefully direct the company's investment in language training, rather than spending its limited budget indiscriminately on all the employees. The same argument can apply to language education in schools and universities, *mutatis mutandis*.

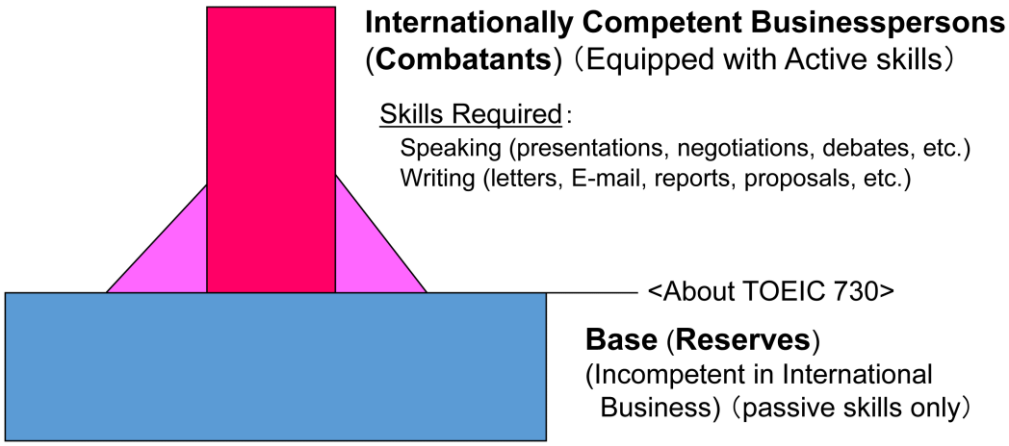

Figure 6. Hierarchical education model of passive base skills followed by active advanced skills.

In this context, it would be worthwhile to review the requirements for language testing from a business perspective. First and foremost, companies should check how well the test they employ aligns with their business objectives and priorities. Without establishing their objectives and priorities and without looking critically at whether the test they employ corresponds with them, it is all too easy to fall victim to a lock-step mentality. No matter how good a test is as a general English test, it is pointless, and therefore a waste of money, to employ it if, for example, what the company needs most is business English competency. In addition to common selection criteria such as validity, reliability, and discriminability, I would list comparability (alignment with an international standard such as the CEFR), ease of use as a management tool, and convenience (ease of administering the test).

Misperceptions and misuses of English tests

In this section I would like to address some of the common misperceptions and the resulting misuses of English tests that are pervasive in Japan. The first misperception is that one test form can uniformly cover all levels – in other words, the notion that if the test can evaluate intermediate levels accurately, then it should also evaluate higher and lower levels equally well. The reality with most tests, however, is that the discriminability is rather poor at the very high and very low ends because of so-called ceiling or boundary effects. Suffice it to say that, in multiple-choice tests for example, one can easily obtain one third or one fourth of the full score by randomly or blindly selecting from the given lists of choices, even without knowing anything at all of the language. It is therefore not ideal to depend on a one-size-fits-all test covering all levels without regard to its intended use.

Another common misperception is to assume that a high score in one test automatically guarantees competence in the workplace. A typical example is the widespread notion that since an employee has scored over 800 in the TOEIC, he/she must be qualified to work as an international businessperson. In actuality, many of those with high TOEIC scores cannot write satisfactory business email or actively participate in business or technical meetings, since there is much more to doing business than merely using the language. Two major factors should be noted. First, the correlation between passive skills and active skills is not high enough to justify such assumptions, and second, language testing and real-life business are different domains. These two points are famously demonstrated in Figures 7 through 9 (Hirai, 2012a; Hirai, 2012b).

Figure 7 illustrates how BULATS Speaking Test scores correlate with TOEIC RL scores. While BULATS speaking scores do tend to increase as TOEIC scores increase, there is a great variance. BULATS Level

3 is the minimum required level for international businesspeople according to a nation-wide survey conducted by Koike, Takada, Matsui, and Terauchi (2010). The regression line shown in red intersects with the BULATS Level 3 line at TOEIC 930. It is also worth noting that 56% of holders of TOEIC 800 or more fail to reach BULATS Level 3. Figure 8 shows the correlation between BULATS Writing Test scores and TOEIC RL scores. The contrast between the two is more dramatic. The regression line does not intersect with the BULATS Level 3 line at all, which means that even with a perfect TOEIC score of 990, more than half of candidates would not reach BULATS Level 3 in business writing. Also, 71% of holders of TOEIC 800 or more fail to reach BULATS Level 3. Figure 9 compares the standard (RL) BULATS test scores of university students and non-student adults. The fairly substantial difference in average score (31.5 vs. 49.6) signifies that in the domain of business English, industry (work) experience plays a significant role.

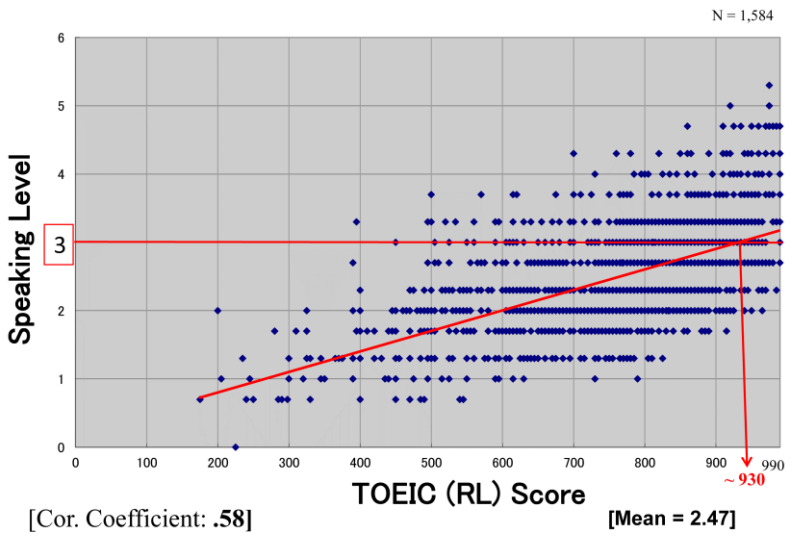

Figure 7. Comparison of BULATS Speaking Test scores with TOEIC Reading and Listening scores.

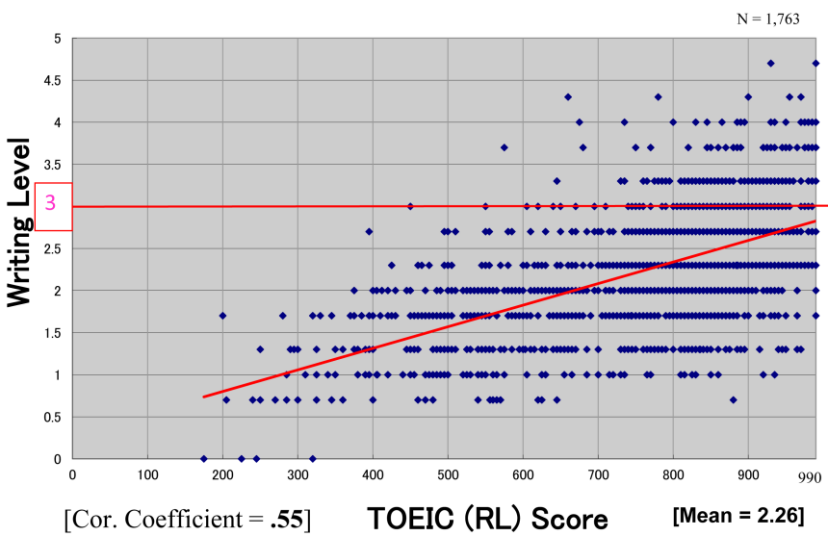

Figure 8. Comparison of BULATS Writing Test scores with TOEIC Reading and Listening scores.

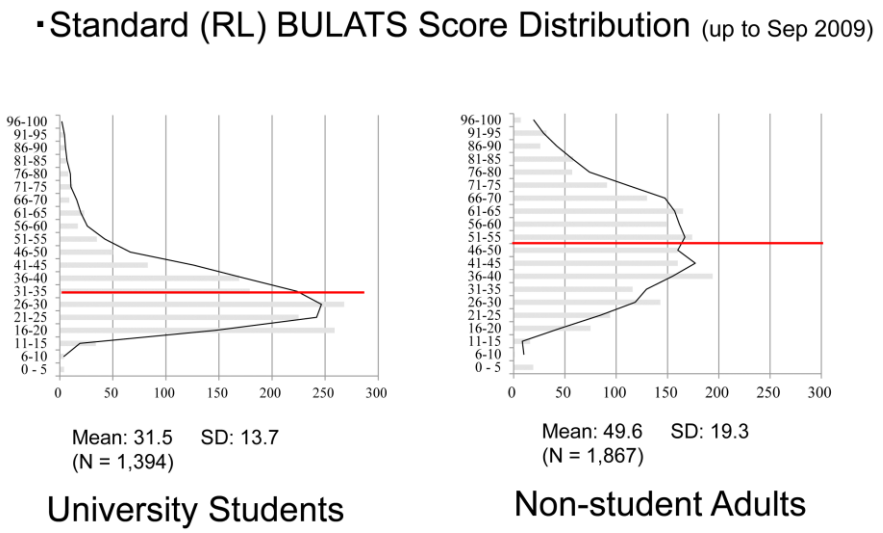

Figure 9. BULATS Reading and Listening test scores for university students and non-student adults.

These findings underscore the importance of choosing the right test for the right purpose. Whereas the TOEIC is very well designed as a general English test (more recently with additional workplace relevance), it would be a big mistake to mindlessly interpret the TOEIC RL score as a reliable indicator of business English competence. A test can only be "good" within the boundaries it is designed for and when its validity (i.e., its ability to test what it claims to test) is warranted. How to use a test is much more important than whether it is a "good" or "bad" test.

Looking ahead

In closing, it would be worthwhile to quickly review what is happening in the field of language testing and try to predict where it is heading. First of all, there is an inexorable shift towards online services. Second, the world is becoming increasingly aware of standardization initiatives such as CEFR, which allows us to compare and evaluate tests and materials with a common measure. Third, test developers continue to improve and enhance their tests by aligning them more closely to how English is used in real life. One example is the proliferation of ESP tests that purport to better serve the needs of industry. In academic English, there is a trend towards the integration of multiple skills. Finally, as with Go and Shogi, artificial intelligence seems likely to eventually make inroads into the world of testing, not to mention the sacred realm of scoring and grading. It may be time for mere humans to pack their bags and retire?

Conclusions

Language testing has many facets and should be viewed from various angles. From the viewpoint of education, it is important to realize how language tests affect the learners' study patterns and hence the formation of their skill profiles. In this regard, four-skills tests are much more desirable (if cost permits) than passive-skills-only tests, because active skills are what is needed most in the real world. From the viewpoint of language-test users, it is essential first to realize the organization's objectives and priorities and then to choose the test (or test battery) that most closely aligns with them. In Japanese industry today, it is crucial to use English tests as a means of fostering well-balanced skill profiles of employees in order to meet their international business requirements. In this respect, it is strongly recommended to use an

English test or test battery that is specifically designed to evaluate the four skills required in actual business situations.

References

Chapman, M. (2003). TOEIC®: Tried but undertested, *Shiken: JALT Testing & Evaluation SIG Newsletter*, *7*(3), 2-7. Retrieved on April 29, 2017 from http://jalt.org/test/cha_1.htm

Cheng, L., & Watanabe, Y. (Eds.). (2004). *Washback in language testing: Research contexts and methods*. Mahwah: Lawrence Erlbaum Associates.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press

ELT Services Japan. (2017). 英語の資格と検定試験, Retrieved on January 9, 2017 from http://www.eigokyoikunews.com/database/exam/

Hirai, M. (2002). Correlations between active skill and passive skill test scores, *Shiken: JALT Testing & Evaluation SIG Newsletter*, 6(3), 2-8. Retrieved on April 29, 2017 from http://jalt.org/test/hir_1.htm

Hirai, M. (2012a). Correlation between BULATS Speaking/Writing and TOEIC Scores. In Chartrand, R., Crofts, S., & Brooks, G. *The 2012 Pan-SIG Proceedings* (pp. 118-125). Tokyo: JALT. Retrieved on April 29, 2017 from http://pansig.org/archive

Hirai, M. (2012b). 受信型スキルテストで仕事における発信型能力を測れるか, *BULATS Journal 2012-2013*, 2-3. Retrieved on April 29, 2017 from http://www.hirai-language.com/wordpress/wp-content/uploads/2013/05/BULATS-Tsushin-2012-2013-Special-Issue-pp.2-3.pdf

Koike, I., Takada, T., Matsui, J., & Terauchi, H. (2010). *企業が求める英語力* (English Abilities Required by Corporations), Tokyo: Asahi Press.

McCrostie, J. (2010). The TOEIC® in Japan: A scandal made in heaven, *Shiken: JALT Testing & Evaluation SIG Newsletter, 14*(1), 2-10. Retrieved on April 29, 2017 from http://jalt.org/test/mcc_1.htm

Appendix

Homepages of English test websites in alphabetical order as of April 29, 2017

American Translators Association (ATA) Certification: http://atanet.org/certification/index.php

BETA (Businessmen's English Test & Appraisal): http://www.ilc-japan.com/tokyo/corporation/gogaku/beta2/bet

BULATS (Business Language Testing Service) English: http://www.cambridgeenglish.org/exams/bulats/

Cambridge English: Key (KET): http://www.cambridgeenglish.org/exams/key/

Cambridge English: Preliminary (PET): http://www.cambridgeenglish.org/exams/preliminary/

Cambridge English: First (FCE): http://www.cambridgeenglish.org/exams/first/

Cambridge English: Advanced (CAE): http://www.cambridgeenglish.org/exams/advanced/

Cambridge English: Proficiency (CPE): http://www.cambridgeenglish.org/exams/proficiency/

Certified Professional Translator Test (JTA 公認翻訳専門職資格試験): http://www.jta-net.or.jp/about_pro_exam.html

Eiken (Practical English) (英検): http://www.eiken.or.jp/eiken/

GRE (Graduate Record Examination): https://www.ets.org/gre

GTEC (Global Test of English Communication): http://www.benesse.co.jp/gtec/

G-TELP (General Tests of English Language Proficiency): http://www.g-telp.jp/english/

Hon'yaku Kentei (翻訳検定): http://www.jtf.jp/jp/license_exam/license.html

IELTS (International English Language Testing System): https://www.ielts.org/

Kokuren Eiken (English Proficiency Test in the Program of the Official Languages Test of the United Nations) (国連英検): http://www.kokureneiken.jp/

Kougyou Eiken (English Technical Writing Test) (工業英検): http://jstc.jp/koeiken/koeiken.html

Nissho Business Eiken (日商ビジネス英検): https://www.kentei.ne.jp/english

Pitman ESOL (English for Speakers of Other Languages): http://anshin-keiri.com/shikaku_02/01_24.html

TEAP (Test of English for Academic Purposes): https://www.eiken.or.jp/teap/

TEP (Waseda-Michigan Technical English Proficiency Test): http://www.teptest.com/outline.html

TOBiS (Test of Business Interpreting Skills): http://www.cais.or.jp/tobis/index.html

TOEFL (Test of English as a Foreign Language): https://www.ets.org/toefl

TOEIC (Test of English for International Communication): http://www.iibc-global.org/english/lr.html

TQE (Translator Qualifying Examination): http://tqe.jp/

Questions and answers about language testing statistics:

Consistency in research design: Categories and subcategories

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

Question:

This column responds to an email I recently received that raised what is clearly the most concise, even terse, question I have ever received for this column: "Hello....what is the exact difference between external reliability and internal reliability in quantitative research?"

Answer:

This is the second of two columns. In both columns, consistency is defined simply as the degree to which something is systematic. I discussed consistency in measurement in the last column as shown in the rectangles to the left with grey backgrounds in Figure 1 in terms of norm-referenced test (NRT) reliability and criterion-referenced test (CRT) dependability. In this column, I will discuss consistency in research design which comes in three flavors: quantitative reliability, qualitative dependability, and mixed methods research (MMR) dependability (see the rectangles to the right with the white backgrounds in Figure 1).

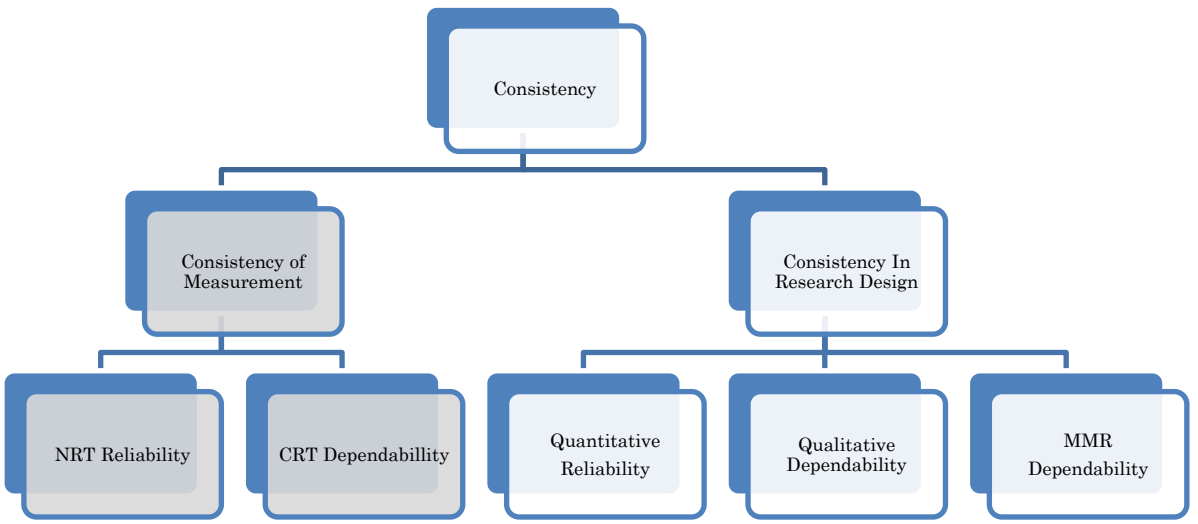

Figure 1. Consistency in measurement (grey) and research design (white)

Consistency in research design categories and substrategies

As mentioned above, consistency in research design falls in three categories: quantitative reliability, qualitative dependability, and MMR dependability, and each of those can be further subdivided into two or three subcategories (as shown in Figure 2).

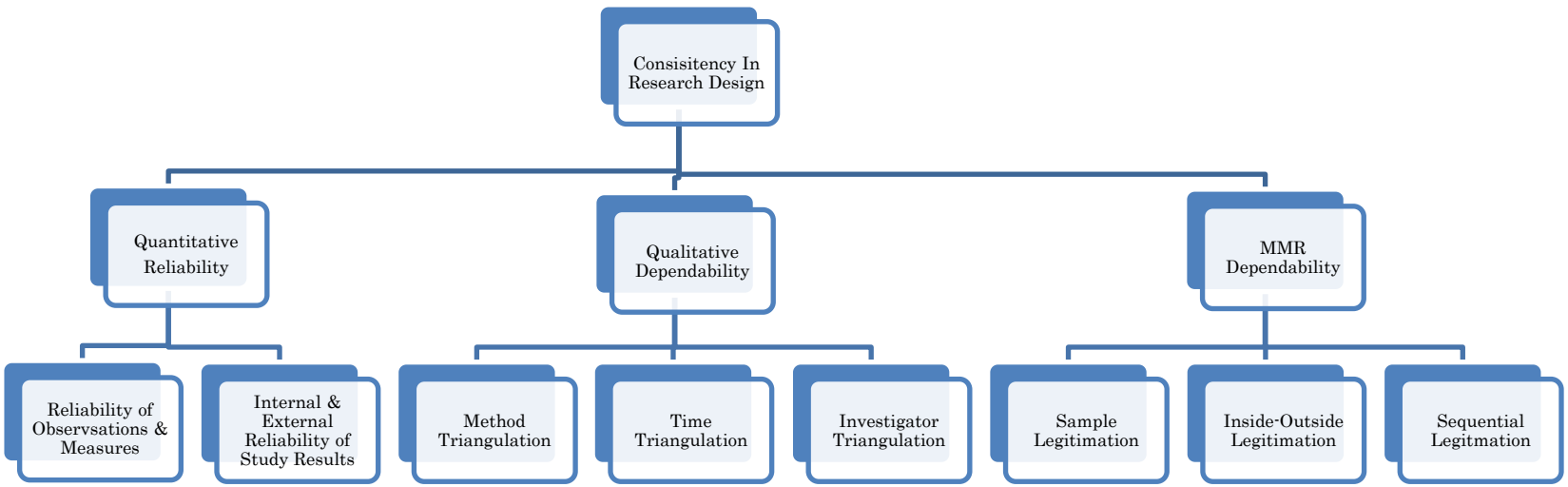

Figure 2. Consistency of in research designs: Categories and subcategories

Quantitative reliability. Quantitative reliability has to do with the degree to which the results of observations/measures are consistent in a study, but also the degree to which the results of the study as a whole are consistent internally and externally. Thus, enhancing or confirming the reliability of a quantitative study should use at least two strategies. The first of these, the reliability of observations/measures strategy can be confirmed or enhanced by calculating reliability estimates for measures or agreement estimates for ratings/codings. For example, for those measures that are based on tests, test–retest, parallel forms, or internal consistency reliability estimates can be calculated, or inter-rater/intra-rater reliability estimates for ratings (for much more on this topic, see Bachman, 2004, pp. 153-91; Brown, 2005, pp. 169-98, 2012), while calculating agreement estimates may be more appropriate for other sorts of observations based on ratings or codings (see Brown, 2001 pp. 231-40).

In contrast, the reliability of the results of the study as a whole can be enhanced *internally* by carefully monitoring and controlling issues that might contribute to inconsistency in design like (a) changes over time due to self-selection of participants into (i.e., using volunteers) or dropping out of the study, (b) maturation in the participants, (c) the Hawthorne effect, (d) the halo effect, or (e) subject/researcher expectancy effects. The reliability of a study can be enhanced/verified *externally* by inspecting the statistical tests that are run in a study with an eye to determining the degree to which the results of the study would be likely to be stable if the study were replicated; for instance, by recognizing that a significant result (at say $p < .01$) means that there is only a 1% chance that the result is due to chance, external reliability can be addressed by thinking about what such probabilities mean for the stability of the results in replication [For definitions and discussion of the terminology used in this paragraph, see Brown, 1988, pp. 29-42, or 2016, pp. 49-53, 162.]

Qualitative dependability. The idea of dependability in qualitative research involves confirming or enhancing the consistency of observations and the effects of changing conditions in the study. Enhancing or confirming the dependability of a qualitative study can use one or more of at least three strategies. The first of these is *method triangulation* (aka overlapping methods), which means using multiple data gathering techniques; for example, a study might include interviews, classroom observations, and a Likert-item questionnaire so that the researcher can examine the dependability of results across methods. The second strategy involves using *time triangulation* (aka stepwise replications), which means gathering data at multiple times; for instance, qualitative data could be gathered at the beginning, middle, and end of a school term so that the dependability of the results over time could be examined. And a third strategy would be to use *investigator triangulation* (aka auditor and inquiry audits), which means having multiple investigators work on the study; for example, qualitative data could be coded by two different investigators with the goal of examining the dependability of codings across investigators. [For more on this terminology and these strategies, see Brown, 2001, pp. 227-231; 2016, p. 158.]

MMR dependability. Since MMR dependability focuses on the consistency of combining quantitative and qualitative data, the consistency of those underlying data and interpretations are a precondition. That is, the reliability of the quantitative data and results should be confirmed or enhanced with regard to the measure/observations and the study as a whole by using the strategies described two subsections above, and the dependability of the qualitative data and results should be confirmed or enhanced with regard to the consistency of observations and effects of changing conditions in the study by using the strategies (i.e., method, time, and investigator triangulation) described in the previous subsection. However, from the additional MMR perspective, the dependability of the efforts to combine quantitative and qualitative data should be examined using at least three types of legitimation: sample, inside-outside, and sequential legitimation. *Sample legitimation* involves examining or enhancing the ways that the qualitative and quantitative samples were integrated and consistent within a study; for example, by examining the consistency of results from qualitative interviews and classroom observations, then examining the quantitative Likert item questionnaires developed from those interviews and observations, and checking

all of that with qualitative follow-up interviews used for member checking. *Inside–outside legitimation* involves considering how adequately the insider (emic) and outsider (etic) perspectives were combined in the quantitative and qualitative data and analyses; for instance, by studying the degree to which the emic perceptions of students and teachers in an institution gathered in qualitative interviews compared or combined with the etic perceptions of the public about that institution gathered in quantitative Likert item questionnaires. *Sequential legitimation* examines the degree to which the effects of method sequencing were minimized; for example, by considering the degree to which results based on interviews conducted before and after the administration of the Likert item questionnaire were consistent. [For more on the concepts discussed in this paragraph, see Brown, 2014, especially pp. 127-135.]

Table 1

Summary of Research Consistency Categories and Subcategories in Quantitative, Qualitative, and MMR Research with Examples

Type	Category	Subcategory	Example
Quantitative Reliability	Reliability of Observations & Measures	Enhanced/confirmed by calculating reliability estimates for measures; or calculating agreement coefficients for ratings or codings	For tests, calculating reliability estimates like test-retest, parallel forms, or internal consistency (e.g., Cronbach alpha, K-R20, etc.) or inter-, or intra-rater reliability estimates; For other sorts of observations, calculating rater/coder agreement coefficients or kappa
	Internal & External Reliability of Study Results	Internal reliability– enhanced/confirmed by controlling issues that often contribute inconsistencies in study design	Monitoring & controlling issues like self-selection, mortality, maturation, Hawthorne effect, halo effect, or subject/researcher expectancies
		External reliability – enhanced/verified by inspecting statistical results in terms of replication	Recognizing that a significant result (at say $p < .01$) means that there is only a 1% chance that the result is due to chance, external reliability can be addressed by thinking about what such probabilities mean for the stability of the results in replication
Qualitative Dependability	Method Triangulation	(aka overlapping methods) Enhanced/confirmed by using multiple data gathering methods	For example, using interviews, classroom observations, & a Likert item questionnaire, & examining dependability of results across methods
	Time Triangulation	(aka stepwise replications) Enhanced/confirmed by gathering data at multiple times	For example, gathering data at beginning, middle, & end of school term, & examining dependability of results over time
	Investigator Triangulation	(aka auditor & inquiry audits) Enhanced/confirmed by using multiple investigators	For example, using two investigators to independently code the data in a study, & examining dependability of results across investigators
MMR Dependability	Sample Legitimation	Enhanced/confirmed by examining how the qualitative & quantitative data samples are integrated & consistent	For example, examining the consistency of results from qualitative interviews & classroom observations, quantitative Likert item questionnaires developed from those interviews & observations, & qualitative interviews used for member checking later in the study
	Inside-Outside Legitimation	Enhanced/confirmed by considering how adequately the insider (emic) & outsider (etic) perspectives were combined in the quantitative & qualitative data & analyses	For instance, by studying the degree to which the emic perceptions of students & teachers in an institution gathered in qualitative interviews compared or combined with the etic perceptions of the public about that institution gathered in quantitative Likert item questionnaires
	Sequential Legitimation	Enhanced/confirmed by examining the degree to which the effects of method sequencing were minimized	For example, by considering the degree to which results based on interviews conducted before & after the administration of the Likert item questionnaire were consistent

Conclusion

In direct answer to your question, "the exact difference between external reliability and internal reliability in quantitative research" is not a very clear, helpful, or adequate way of characterizing the consistency issues that arise in of consistency of measurement or consistency of research design.

In the previous column, I addressed the issues involved in consistency of measurement separately for norm-referenced and criterion-referenced tests. In the present column, I have shown that consistency in research design, comes in three categories: quantitative reliability (with subcategories for consistency of observations and measurement and consistency of study results internally and externally), qualitative dependability (with subcategories for method, time, and investigator triangulation), and MMR dependability (with subcategories for sample, inside-outside, and sequential legitimation). Table 1 summarizes all of those aspects of consistency in research.

I hope this and the preceding column together have addressed your question and helped you to realize that a simple external/internal reliability categorization of the issues involved is neither complete nor useful.

References

Bachman, L. F. (2004). *Statistical analysis for language assessment.* Cambridge: Cambridge University.

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design.* Cambridge: Cambridge University.

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge: Cambridge University Press.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment (New Edition)*. New York: McGraw-Hill.

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

Brown, J. D. (2016). *Statistics corner: Questions and answers about testing statistics.* Tokyo: Testing and Evaluation Special Interest Group of JALT.

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 December 2017 issue. Submissions received by 1 September, 2017 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.*