SRB
SHIKEN RESEARCH BULLETIN
Volume 17 • Number 1 • May 2013

Contents

JALT
Testing and Evaluation SIG

Foreword

Jeffrey Stewart
TEVAL SIG Publications Chair,
Shiken Research Bulletin *General Editor*

As a long-time reader, there were three things in particular that I liked about *Shiken Research Bulletin* before joining the editorial board. One was the ability of its contributors to use modern test theory and other statistical approaches to shed new light on issues that we commonly face as language testers. Second was the ability of its contributors, particularly long-time columnists JD Brown and James Sick, to explain how to use these same statistical approaches in a manner that is accessible to readers. Third were the interviews with luminaries in our profession, which shed light not only into the important issues of today, but also into the lives and careers of the people that make our field so vibrant.

These fine attributes are in full display in our most recent issue. In regards to examining issues in our field with modern methods, Judith Runnels employs a polytomous Rasch model to examine the rank order of the developmental scale implied by the CEFR-J, an increasingly widespread version of the CEFR intended for Japanese learning contexts. Our Editor-at-Large Aaron Batty carefully details how a model by Paek et al. can allow us to detect and compare important discrepancies between learners' subjective judgments of their own knowledge and the objective outcomes revealed under testing. In regards to education, our statistics columnist JD Brown explains not only how readers can easily conduct chi-square tests with free online software, but also explains common situations in which the analysis is *not* appropriate, and guides readers towards alternative, but equally simple, approaches. In turn, our Rasch columnist Jim Sick addresses the issue of disjoint subsets in FACETS which leads to discussion of judging plans, a matter of critical concern to any language tester organizing a speaking test with multiple judges. Finally, part II of our interview with Alistair Van Moere, which focuses on the Versant Spoken Language Test developer's career history, will be of particular interest to young researchers who have also began their careers in language testing in Japan, and are interested in beginning PhDs and embracing the full range of exciting possibilities in language testing. I hope you will enjoy this issue as much as I have.

Preliminary validation of the A1 and A2 sub-levels of the CEFR-J

Judith Runnels
jrunnels@h-bunkyo.ac.jp
Hiroshima Bunkyo Women's University

Abstract

The newly released Common European Framework of Reference Japan (CEFR-J) was designed to address the issue that a consistent system for measuring learner proficiency and progress in foreign language pedagogy in Japan is lacking. This tailored version of the Common Europe Framework of Reference (CEFR) was developed to better discriminate incremental differences in proficiency for Japanese learners of English, who tend to fall mostly within the A1 and A2 levels. Changes from the original CEFR included the creation of can-do illustrative descriptors that separated 4 of the existing 6 levels into sub-levels. The goal of the current analysis is to test the suitability of the new sub-levels of A1 and A2 for target users of the system in two ways: 1) by determining if newly developed descriptors are empirically rank ordered by difficulty as specified by the CEFR-J, and 2) by testing the statistical significance of differences in difficulty ratings between the sub-levels. The current analysis found that the rank ordering of levels was the same as predicted by the CEFR-J, and that the higher-order A1 and A2 levels varied in difficulty to a statistically significant degree, but significant differences between adjacent CEFR-J sub-levels were not found. This raises questions about how users of the system can effectively distinguish features representative of each level and whether the additional sub-levels in the CEFR-J can function as intended. Limitations of using a system of illustrative descriptors based primarily on estimates of difficulty and the process of contextualizing a generalized framework are discussed.

Keywords: Common European Framework of Reference, CEFR-J, can-do statements, difficulty, contextualization

Theoretical work, case-studies and other evidence have suggested that the Common European Framework of Reference (CEFR) provides an effective scheme for describing the needs and outcomes of study for language learners (Council of Europe, 2001). The CEFR "describes in a comprehensive way what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively. The Framework also defines levels of proficiency which allow learners' progress to be measured at each stage of learning" (Council of Europe, 2001, p. 1). Some argue that the CEFR "is now accepted as the international standard for language teaching and learning" (North, Ortega, & Sheehan, 2010, p. 6).

The framework operates via illustrative descriptors, often referred to as can-do statements, that act to describe what learners are capable of at any given point in time (North, 2007). The CEFR's can-do statements were developed using both qualitative and quantitative methodologies for each level (North, 2000; North & Schneider, 1998). They represent a communicative scheme that gradually progresses from easy to more difficult and are worded in positive terms (Trim, 1978), such that each statement provides a self-sufficient criterion which allows it to be defined independently from other descriptors (Skehan, 1984). The can-do statements are divided into six proficiency levels ranging from Basic User (levels A1 and A2), Independent User (B1 and B2), to Proficient User (C1 and C2) for five skills (listening, reading, spoken production, spoken interaction and writing). Their ultimate aim is to provide a set of learner-centered, performance-related scales which allow for standardized assessment of level (North, 2007).

However, it is in the area of level assessment that the CEFR's suitability and usefulness are frequently questioned and most heavily criticized, particularly with regards to how the can-do statements should be used for test design and evaluation (Weir, 2005). Weir (2005) cautions that in its current form, the CEFR is neither "comprehensive, coherent or transparent for uncritical use in language testing" (p. 281). One issue is related to defining what is entailed by the notion of can-do mastery, a conceptual problem

that exists for any system employing illustrative descriptors. As North (2007, p. 13) writes, "what exactly do we mean by 'can do'? Should it be certain that the person will always succeed perfectly on the task? This would be too stringent a requirement. On the other hand, a 50 per cent chance of succeeding would be too low to count as mastery." In order for a system of descriptors to be effectively useable, a definition of mastery is required that is described in terms of how likely it is that a person at a certain level can succeed at a task specified by the can-do - this is one aspect that is sorely lacking from the CEFR. Further criticisms relate to the absence of any theoretical basis in or demonstrable link to work in second-language acquisition (Hulstijn, 2007) when other frameworks, such as the guidelines suggested by the American Council on the Teaching of Foreign Languages (ACTFL) have made this a priority for the last few decades (see Pienemann, Johnston, & Brindley, 1988). Furthermore, there is growing evidence that uninformed usage of the CEFR is leading to assumptions that the CEFR's scales directly tie to stages of language acquisition or specific levels in tests such as the Test of English for International Communication (TOEIC) when in fact, it is derived primarily from difficulty judgments made by language educators (Fulcher, 2003, 2010). Additionally, there is little to no evidence to support the CEFR's pedagogic arguments for gradual development across levels: the hierarchy of difficulty and uni-dimensional or linear progression from easy to more difficult entailed by the framework remains largely unsupported by empirical evidence of performance samples (Westhoff, 2007). Finally, as a basis for test development or any other measures of proficiency, the CEFR and its derivatives do not provide sufficient guidelines for the development of standardized assessments since they lack the information required for either generating test specifications, or being "a medium by which existing tests and specifications can be compared" (Fulcher, 2010, p. 19; see also Fulcher, 2004). Even North (2002), a coauthor of the framework, warns against its usage without full comprehension of its limitations. Ultimately, it is frequently noted by supporters and non-supporters alike, that as long as the CEFR is seen only as a heuristic model and that its limitations are kept in mind, it can nonetheless be employed as a practical and useful tool in constructing curricula, materials and assessments (Fulcher, 2010; North & Schneider, 1998).

In Japan in particular, there is currently no consistently used system for the measurement of achievement of English language learners. Negishi (2011) describes an urgent need for the introduction of a common language framework in Japan in order to start moving towards the widespread, consistent usage of a standardized system for foreign language learning, teaching and assessment. Others have argued the benefits of applying such a system, and specifically the CEFR, to pedagogy in Japan (O'Dwyer & Nagai, 2011). The CEFR was selected as a suitable outlet and serious research on the implementation of the CEFR to foreign language pedagogy in Japan began in 2008 at the Tokyo University of Foreign Studies (Negishi, Takada & Tono, 2011). Difficulty surveys of can-do statements from the DIALANG self-assessment statements (Council of Europe, 2001, pp. 231-234) were administered to 360 Japanese university students. Since they ordered consistently with the CEFR's rank ordering of difficulty, it was concluded that the system was applicable to Japanese learners. Additional findings also demonstrated that over 80% of language learners in Japan skewed towards the A and B levels of the scale (Negishi, 2011). It was concluded that the can-dos across these two levels neither effectively distinguished nor adequately accounted for the variation of ability of language users and development of an alternative version was announced (Negishi, 2011).

Known as the Common European Framework of Reference Japan (CEFR-J), this new version encompasses the following modifications from the CEFR (Tono & Negishi, 2012):

- addition of a Pre-A1 level
- division of A1 into three levels: A1.1, A1.2, A1.3
- division of A2 into two levels: A2.1, A2.2
- division of B1 into two levels: B1.1, B1.2
- division of B2 into two levels: B2.1, B2.2
- adapted can-do statements to a Japanese context

There is a dire need for empirical support of these new statements and level divisions prior to widespread implementation in pedagogy in Japan. As there is currently little research to draw upon from within a Japanese context, the current study was designed to establish a starting point for further research on the newly developed level divisions, or sub-levels, of the CEFR-J at the A1 and A2 levels. Using fabricated or contextualized can-do descriptors has been argued to raise a fundamental question of validity: can a framework function both as a generic reference point and also as a specific application in a local context (North, 2007)? In other words, does the CEFR, which was largely developed and researched within a European context, remain a useable pedagogical tool following modifications and application to a Japanese context? The current study will address this question in two ways: 1) by testing the rank ordering of the can-do statements to determine consistency with the CEFR-J, and 2) by determining if the difference in difficulties between the sub-divisions of levels and categorization of can-dos into each level are statistically significant. Since the CEFR illustrative descriptors have empirically supported interpretations of difficulty (represented in their levels), these difficulty levels should remain consistent if the system is to remain applicable to language regions or educational sectors differing in circumstances to the initial location of development (North, 2007). The first hypothesis is therefore that participants of the current study (target users of the system) will order the can-do statements in the same way as specified by the CEFR-J. Disordered levels would represent a lack of the progression of difficulty entailed by the levels of the CEFR-J and question the underlying assumptions of the system. Secondly, since production of a scale is only the first step in the implementation of a framework, ensuring a common interpretation through empirical support is necessary (North & Schneider, 1998). This requires the existence and identification of features which distinguish one level from the next, or in other words, differences between the estimated difficulties of the newly developed levels. The second hypothesis is therefore that the measures of difficulty across sub-levels will differ significantly from each other. Lack of differences in difficulty would question the thresholds of performance or ability or the features of language required for distinguishing between levels and could result in inconsistent judgements of proficiency.

Methods

Participants

296 first and 294 second-year students of Hiroshima Bunkyo Women's University participated voluntarily in this study. Participants were in one of five disciplines of study: Early Childhood Education, Welfare, Nutrition, Psychology and Global Communication. The survey was administered at the end of July 2012, meaning that the former four major students had completed at least one or three semesters of twice weekly 90 minute university level English classes. The Global Communication majors (a total of 12.5% of participants) had completed either one or three semesters of full-time English study.

Instrument

The survey was administered online using www.surveymonkey.com (SurveyMonkey.com, 2012). Participants were required to indicate the extent of their agreement on a 5-point Likert scale to all 50 Japanese can-do statements for all five skills from levels A1.1 to A2.2. The statements were presented in a random order. These levels were selected because they are the target levels for the institution's curriculum.

Procedure

Since each CEFR-J level is divided into two statements for each of the five skills, there are 10 statements for each level. Due to this being a preliminary investigation, the mean difficulty was calculated for all statements across all skills for each level. The Rasch-measurement software package Winsteps (Linacre, 2010) and PASW Statistics 18 were employed for analysis. The mean Rasch measure, in logits, was calculated for each of the CEFR-J levels from A1.1 to A2.2. Difficulty comparisons across levels were carried out in two ways: first, by measuring differences in the mean logit ratings for each level (where a logit difference of 0.3 represents a significant main effect for difficulty; Lange, Greyson, Houran, 2004; Miller, Rotou, & Twing, 2004) and second, with an ANOVA followed by a least significant difference (LSD) post-hoc test.

Results

Descriptive statistics for the items are shown in Table 1, where a lower logit score represents a lower rating of difficulty. It can be seen that A1.1 had the lowest difficulty rating and A2.2 had the highest, with the remaining levels proceeding in ascending order. The item sub-levels did indeed order by difficulty as hypothesized, although the mean difficulties for each level were very close to each other. This is also evident in Figure 1, which shows a Rasch pathway for the CEFR-J levels. In Figure 1, each level is represented with a circle, whose size is proportional to the standard deviation of the measure for that level. Infit-mean squares are shown on the x-axis.

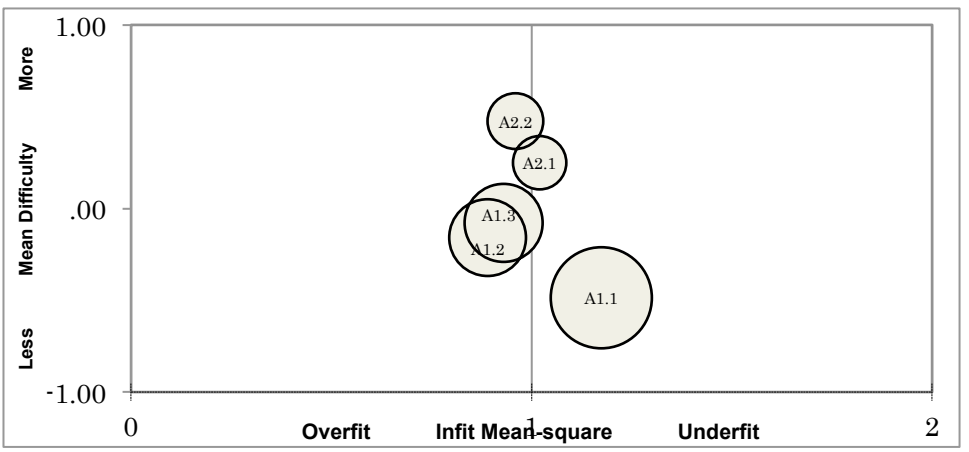

Figure 1. **The bubble chart for CEFR-J levels A1.1 to A2.2.**

Table 1. ***Descriptive Statistics for CEFR-J Levels***

CEFR-J Level	Mean Difficulty	S.D.
A1.1	-0.49	0.584
A1.2	-0.16	0.443
A1.3	-0.08	0.451
A2.1	0.25	0.308
A2.2	0.48	0.322

None of the items exhibited mean-squares outside of the 0.7 – 1.2 range deemed acceptable by Wright and Linacre (1994) and fit statistics for the items have therefore been omitted. As shown in Table 1, the difference in difficulty between levels exceeds the 0.3 logit difference required for significance, for levels A1.1 and A1.2 (0.33) and between A1.3 and A2.1 (0.33). The required difference of 0.3 logits for significance between levels A1.2 and A1.3, or A2.1 and A2.2 was not found. An ANOVA was performed to examine the relationships between levels in more detail. Although differences in difficulties between levels were significant overall ($p = <0.001$; $R^2 = 0.396$), an LSD post-hoc test revealed that there were no significant differences between any adjacent sub-levels (Table 2).

Table 2. ***LSD Post-Hoc Tests for Adjacent Categories***

		Mean Difference	Std. Error	Sig.
A1.1	**A1.2**	-0.33	0.194	0.096
A1.2	**A1.1**	0.33	0.194	0.096
	A1.3	-0.08	0.194	0.693
A1.3	**A1.2**	0.08	0.194	0.693
	A2.1	-0.33	0.194	0.098
A2.1	**A1.3**	0.33	0.194	0.098
	A2.2	-0.23	0.194	0.246
A2.2	**A2.1**	0.23	0.194	0.246

Interestingly, when items were grouped by the original A1 and A2 categories of the CEFR itself, rather than using the sub-levels of the CEFR-J, a statistically significant difference was found. In this case, the overall A1 mean difficulty was -0.24 and the overall A2 mean was 0.36, for a difference of 0.6 logits ($t = 5.075$; $p = <0.001$).

Discussion

The analyses herein were designed to provide empirical evidence on the difficulty of the newly developed levels of the CEFR-J. The first hypothesis tested the rank ordering of difficulty of CEFR-J level statements. The results indicated that the participants ranked the difficulties of the sub-levels in the same way as specified by the CEFR-J (Figure 1 and Table 1). This is not surprising given the extensive process undertaken to create the CEFR-J's can-do descriptors (see Negishi, 2011). Furthermore, previous studies have demonstrated that high correlations between the rank ordering of the difficulty of fabricated descriptors are common (see Jones, 2002; Kaftandjieva & Takala, 2002). Nonetheless, this finding is only preliminary as it compared solely the overall mean difficulty of the sub-levels.

The second hypothesis that differences in difficulty across levels would exist was, unlike the first hypothesis, not supported. Not only did levels A1.2 and A1.3, as well as A2.1 and A2.2 lack the logit

difference of 0.3 considered necessary for a main effect of difficulty, but the differences between the remaining levels (A1.1 and A1.2; A1.3 and A2.1), only just meet the threshold of 0.3 logits. When more specific testing was performed using ANOVAs, no significant differences were found between adjacent CEFR-J levels. This raises questions about the division of level A1 and A2 into three and two sub-levels respectively, since the ratings made by users of the system indicate that there seems to be very little to distinguish features representative of these divisions. After reverting back to the divisions of the original CEFR, however, the difference in mean difficulty between the higher order levels of A1 and A2 was both larger and statistically significant. This suggests that the proposed sub-levels for A1 and A2 in the CEFR-J may attempt to make a finer distinction in proficiency than is realistically possible. This will likely represent a challenge for those attempting to place users within the A1 and A2 range, which is also where the majority of Japanese users are purported to lie (Negishi, 2011). As is discussed in Council of Europe (2001, p. 21): "the number of levels adopted should be adequate to show progression…but should not exceed the number of levels between which people are capable of making reasonably consistent distinctions." A potential solution may be to reduce the A1 sub-divisions from three to two and perhaps even A2 into a single level. The same situation may also exist for the sub-divisions of the B1 and B2 levels: further research on this is required to determine if this is a possibility. In either case, this relates back to criticisms of the CEFR: that the assumptions inherent in the hierarchy of levels require supporting empirical evidence (Westhoff, 2007).

A major drawback to the current results however, is that difficulty ratings were averaged across the entire CEFR-J level such that the difficulty was not broken down into separate skills. The data presented herein represent the mean for the entire level across all of the five skills. Future studies should aim to measure the equivalencies between the two can-do statements for each skill for each level, and also across the separate skills. Doing so would better ensure a gradual progression of difficulty across the levels,.

Further limitations of the current study relate to the usage of self-assessment data. It is possible that participants' estimations of whether they have mastered material implicated by the can-do statements are inaccurate: no controls for ability have been employed herein (although this possibility also exists for the participants who were involved in the creation of the system initially—see Negishi, 2011). Investigating this would require comparisons of ability or proficiency derived through other forms of assessment to ensure that more abled students are agreeing with their achievement of the can-dos at higher rates than their lesser abled counterparts. Indeed, one of the criticisms of the CEFR is hinged on this same aspect: that while self-assessment by a language learner or assessment by a language teacher produces scales which order consistently in difficulty between these groups, these scales lack reification (Fulcher, 2010). In other words, it is insufficient for a language framework, if it is to be called that, to be solely based in difficulty estimations by its users (either students and/or teachers), particularly if, as the results in the existing study seem to suggest, the users' behavior does not consistently match predictions by the system.

Conclusion

The results indicated that the participants ranked the difficulties of the sub-levels in the same way as specified by the CEFR-J, though in many cases differences in difficulty between adjacent sub-levels were negligible, and below the threshold of 0.3 logits considered to represent a main effect of difficulty (Lange, Greyson, Houran, 2004; Miller, Rotou, & Twing, 2004).

As Trim (1996) notes, the CEFR deliberately lacks details for local decision-making and action. While it can certainly guide characterizations of language use and language pedagogy, it should not be employed or interpreted as a standardized, benchmarked system. As Davies (2008) points out, when large-scale

operations are perceived as 'the system', this has historically resulted in a reduction of diversity and experimentation in research surrounding language pedagogy. Ultimately, the investigation of the process of contextualization of the general CEFR to the more local CEFR-J, has revealed that the preliminary work by Negishi, Takada, & Tono, (2011) was relatively successful. The levels ordered as specified and minor differences between some sub-levels were found when target users of the system provided difficulty ratings, although more evidence is required. The results of the present analysis are only a starting point for further validation studies of the CEFR-J, as they do not make any measurements across the language skills or statements contained at each level nor do they provide any controls for proficiency. Establishing the validity of a specialized system which has been developed from a generic reference point is a challenging endeavor (North, 2007). Development alone does not ensure an effective system of measurement or assessment that is capable of specifying the needs, materials or outcomes of study. Research on the CEFR has spanned over twenty years and is ongoing, with continual updates and modifications: the same is required for the CEFR-J. For quality assurance, the system needs to be subject to empirical testing for applicability and effectiveness at every level prior to full implementation or widespread usage in foreign language pedagogy of institutions in Japan.

References

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

Davies A. (2008). Ethics and professionalism. In E. Shohamy (Ed.), *Language testing and assessment.* (pp. 429-443). New York: Springer.

Fulcher G. (2003). *Testing second language speaking.* London: Longman/Pearson.

Fulcher G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly, 1*(4), 253-266.

Fulcher, G. (2010). The reification of the Common European Framework of Reference (CEFR) and effect-driven testing. *Advances in Research on Language Acquisition and Teaching: Selected Papers,* 15-26.

Hulstijn J. A. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal, 91*(4), 663-667.

Jones, N. (2002). Relating the ALTE Framework to the Common European Framework of Reference. In J. C. A. Alderson (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment. Case studies.* (pp. 167-183). Strasbourg: Council of Europe.

Kaftandjieva, F., & Takala, S. (2002). Council of Europe Scales of Language Proficiency: A validation study. In J. C. A. Alderson (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment. Case studies.* (pp. 106-129). Strasbourg: Council of Europe.

Lange, R., Greyson, B., & Houran, J. (2004). A Rasch scaling validation of a 'core' near-death experience. *British Journal of Psychology, 95,* 161–177.

Linacre, J. M. (2010). Winsteps® (Version 3.70.0) [Computer Software]. Beaverton, Oregon: Winsteps.com.

Miller, G. E., Rotou, O., & Twing, J. S. (2004). Evaluation of the .3 logits screening criterion in common item equating. *Journal of Applied Measurement, 5*(2), 172-177.

Negishi, M. (2011). CEFR-J Kaihatsu no Keii [The Development Process of the CEFR-J]. *ARCLE Review, 5*(3), 37-52.

Negishi, M., Takada, T., & Tono, Y. (2011). A progress report on the development of the CEFR-J. *Association of Language Testers in Europe Conference.* Retrieved August 1st from: http://www.alte.org/2011/presentations/pdf/negishi.pdf.

North, B. (2000). *The development of a common framework scale of language proficiency.* New York: Peter Lang.

North, B. (2002). Developing descriptor scales of language proficiency for the CEF common reference levels. In J. C. A. Alderson (Ed.), *Common European Framework of Reference for Languages: learning, teaching, assessment. Case studies.* (pp. 87-105). Strasbourg: Council of Europe.

North, B. (2007). The CEFR Common Reference Levels: Validated reference points and local strategies. *Language Policy Forum Report,* 19-29.

North, B., Ortega, A., & Sheehan, S. (2010). A core inventory for general English, British Council/EAQUALS. Retrieved August 3rd from: http://www.teachingenglish.org.uk/publications/british-council-eaquals-core-inventory-general-english.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217–262.

O'Dwyer, F., & Nagai, N. (2011). The actual and potential impacts of the CEFR on language education in Japan. *Synergies Europe, 6,* 141-152.

Pienemann, M., Johnston, M., & Brindley, G. (1988). Constructing an acquisition-based procedure for second language assessment. *SSLA,* 10, 217-243.

Skehan, P. (1984). Issues in the testing of English for specific purposes. *Language Testing, 1*(2), 202–220.

SurveyMonkey.com. (2012). SurveyMonkey. From http://www.surveymonkey.com/

Tono, Y., & Negishi, M. (2012). The CEFR-J: Adapting the CEFR for English language teaching in Japan. *Framework & Language Portfolio SIG Newsletter, 8,* 5-12.

Trim, J. L. M. (1978). Some possible lines of development of an overall structure for a European unit/credit scheme for foreign language learning by adults. In J. C. A. Alderson (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment. Case studies.* Strasbourg: Council of Europe, Appendix B.

Trim J. L. M. (1997). The proposed Common European Framework for the description of language learning, teaching and assessment. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma, (Eds), *Current developments and alternatives in language assessment. Proceedings of the LTRC,* (pp. 415-421). Jyvaskyla: University of Jyvaskyla Press.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Hampshire, UK: Palgrave-Macmillan.

Westhoff, G. (2007). Challenges and opportunities of the CEFR for reimagining foreign language pedagogy. *The Modern Language Journal, 91*(4), 676 – 679.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Applying the Paek et al. method for calculating over- and under-confidence at the item and test levels

Aaron Olaf Batty
abatty@sfc.keio.ac.jp
Keio University
Lancaster University

Abstract

This article explains investigating over- and under-confidence on tests and test items using the method developed by Paek et al. (2008) for use with Rasch and other IRT measures. The data for this demonstration originated from a study of 199 Japanese high school and university students, investigating their knowledge of a number of special uses of verbs of utterance in English. The paper provides practical information on the calculations necessary for the use of the Paek et al. method under the Rasch model and the interpretation of the results. Finally, the same data are scaled with the two-parameter IRT model and the Paek et al. method is applied for comparison.

Keywords: confidence, accuracy, discrepancy, overconfidence, underconfidence, Rasch, two-parameter IRT

Many researchers have attempted to enhance their test data by incorporating a confidence scale to their instruments. Such data can prove useful within an instructional context, as discrepancy between confidence and accuracy signals a need for further instruction at a finer level of detail than simple accuracy (i.e., right/wrong) data can. On such instruments, the respondent typically responds to answers an item and then indicates his or her confidence of the accuracy of the answer on a separate scale. Such instruments are by no means new, and methods for interpreting their results have been in use for decades (e.g. Hakstian & Kansup, 1975; Keren, 1991; Lichtenstein, Fischhoff, & Phillips, 1981), but a method for incorporating confidence data with IRT measures was not developed until more recently.

Paek et al. (2008) describe a method of comparing the accuracy and the confidence of respondents on items and on tests overall, taking advantage of Rasch (1960) IRT measures to control for overall accuracy. It allows the researcher to examine the amount of over- or under-confidence on each item, and to compare it to overall ability on the measure. The method is described in detail in their 2008 ETS research report, but the discussion is not very accessible to most readers in the field of language testing. In this brief article, I will explain the process in greater detail, in the hopes that more language testing researchers can add it to their repertoire of test analyses.

Although a Rasch program is necessary to obtain the difficulty/ability estimates, the remainder of the calculations can be done in a spreadsheet application such as Microsoft Excel. A high-quality graphing software package is necessary to produce the plots discussed at the end of the article.

Data

The data used for illustration of this method are from the continuation of the utterance verb study described in Sato and Batty (2012). The participants were 199 Japanese learners of English, ranging from high school through undergraduate university, with abilities ranging from beginner through bilingual. The items were 20 gap-fill sentences, for which the respondent must choose an appropriate verb of utterance (*speak, talk, say,* or *tell*) and then report his/her confidence of the correctness of his/her answer on a scale from 1 (not confident) through 3 (very confident). Hence, every item has both a dichotomous accuracy variable as well as a polytomous confidence variable.

Method

The method compares values that Paek et al. refer to as the *objective probability* and the *subjective probability*. The former is simply the probability of a given respondent answering a given item correctly, given his/her level, while the latter is the probable confidence level of a person of that level, on that item. The former is conceptualized as a measure of *accuracy*, and the latter of *confidence.*

First, a metric of over/underconfidence must be obtained for each person. This is calculated as the difference between an individual's subjective probability and objective probability, where a positive value (indicating that the examinee's subjective probability is larger than his/her objective probability—or, to put it another way, his/her confidence outstripped his/her ability) is interpreted as overconfidence, and a negative as underconfidence, as below (Paek et al., 2008, p. 3):

$$\text{Overconfidence} = P_i^* - P_i > 0 \tag{1}$$

$$\text{Underconfidence} = P_i^* - P_i < 0 \tag{2}$$

where P_i is the objective probability of item i (whatever it may be in this case, e.g., the objective probability for a certain person of a certain ability on item 15 would be P_{15}) and P^*_i is the subjective probability of item i. The next section explains how to obtain these values.

Obtaining the objective probability scores (P_i)

The first step is to run Rasch estimation on the test. All Rasch packages provide an estimate of the probability of success for each person on each item, although most packages call this statistic *expected score.* It is critical to note, that this is not the probability of the respondent's actual response to the item; it is the probability of the respondent answering the item in question correctly, given his/her overall score and the difficulty of the item.

Following are brief instructions for obtaining the expected score statistic in two popular Rasch estimation packages.

Winsteps

In Winsteps (Linacre, 2012a), expected scores are found in the observation file (aka XFILE) available as an output from the "Output Files" menu. The output will have one line per person, per item. Therefore, for example, if your test has 10 items and you had 25 respondents, there would be 250 lines.

The probability of a correct response by the person on the item is found in the "EXPECTED" column. This is the objective probability for calculating over/underconfidence.

Facets

In Facets (Linacre, 2012b), expected scores are found in the "Residuals/Responses file", available from the "Output Files" menu. The expected scores are found in the column labeled "Exp".

Obtaining the subjective probability scores (P^*_i)

The subjective probability scores are simply the average confidence scores for all the examinees at a certain Rasch ability level, on each item, rescaled to have a range of 0 to 1 (i.e., the range of a probability, regardless of how long the confidence scale on the test in question was).

For example, in the example utterance verb data, the average confidence score on item 1, for all the respondents with a Rasch ability score (denoted θ in mathematical notation; here it is denoted θ_{ac} because it is the ability score on the *accuracy* dimension) of 0.52 is 1.62. This is divided by 3 to rescale it to a number from 0 to 1, representing a probability: 0.54. This rescaled value is the subjective probability for item 1 for every person with a Rasch ability score of 0.52. Because Rasch ability scores are scaled from total raw scores, the number of discrete Rasch scores in any dataset is finite, and, furthermore, is rather small. In the case of the example instrument with 20 items, there are only 21 possible Rasch scores (i.e., one for each number of items correct, from 0 through 20). In practice, however, there were only 18 discrete Rasch ability scores for the respondents.

There are many ways to quickly calculate the subjective probability, but the way I approached it was by using using Excel's VLOOKUP formula to match Rasch scores from the score file to respondents in a worksheet with their objective probabilities. I then used VLOOKUP on another column to bring in the confidence score for each item for each person. Finally, I used the AVERAGEIFS function to average the confidence scores for each item, for each Rasch score (θ_{ac}).

To calculate the subjective probability for the first respondent on the first item in Figure 1, for example, I used the following Excel formula:

```
=AVERAGEIFS(D:D,C:C,C2,B:B,B2)/3
```

which instructs Excel to average values from the D column if the value in the C column ("Ability", or the Rasch ability score for that respondent) matched the value for this person (C2), and if the value in the B column (the item number) matched the value for this particular row (B2). Finally, it divides that average by 3 to rescale it to 0 – 1. This results in a rescaled average of all the confidence scores for people with the same ability score as respondent W007 (i.e., 0.52) on the item in question (item 1).

	A	B	C	D	E	F
1	Examinee	Item#	Ability	Confidence	Obj Prob	Sbj Prob
2	W007	1	0.52	2	0.605	0.540
3	W007	2	0.52	2	0.340	0.575
4	W007	3	0.52	2	0.260	0.632
5	W007	4	0.52	2	0.775	0.621
6	W007	5	0.52	3	0.845	0.713

Figure 1. Calculating the subjective probability.

Calculating the over/underconfidence scores

To calculate the over/underconfidence score for each person and each item, simply subtract the objective probability (aka the Rasch expected score) from the subjective probabilities. Note, however, that the resulting over/underconfidence score will be the same for all people who share the same Rasch ability score (θ_{ac}), on each item. From this point on, the Paek et al. method does not concern itself with individual examinees. For this reason, it is advisable to paste the data into a new sheet and remove the duplicates so that what remains is one list of all of the numbers for each Rasch ability estimate level.

Calculating the Item Discrepancy Index (*IDI*)

The IDI is a summary statistic for the amount and direction of the discrepancy between accuracy (the objective probability) and confidence (the subjective probability), also called the over/underconfidence, on each item of the instrument. It is weighted by the ratio of examinees at each Rasch ability level. This controls for the distribution of ability levels by making more-common ability levels contribute more than less-common ones. It is calculated as below (Paek et al., 2008, p. 4):

$$(IDI_i) = \sum_{\theta_{ac}} (P_i^* - P_i) w(\theta_{ac}) \tag{3}$$

where:

$$w(\theta_{ac}) = \frac{N_{\theta_{ac}}}{N} \tag{4}$$

Calculating the weighting statistics

The first half of equation 3 has already been calculated at this point, as it is simply the over/underconfidence statistic described above. The next calculation is the weighting variable described in equation 4, which is calculated by doing nothing more that counting the number of people at a certain Rasch ability score level (θ_{ac}), and dividing it by the total N of the sample. In a spreadsheet program, this can be accomplished by using the COUNTIF and COUNT functions. The example in Figure 2 uses the following formula to calculate the weighting statistic in the cell F2:

```
=COUNTIF(Data!C:C,'IDI Calcs'!B2)/COUNT(Data!C:C)
```

It first counts up all the rows in the sheet with the full dataset that have the same ability score (θ_{ac}) as appears in B2. It then divides that count by the count of all the rows in the full dataset. Because each person has exactly 20 rows of data, it does not matter that we are counting items, rather than people. The result will be the same.

	A	B	C	D	E	F
1	Item#	Ability	Obj Prob	Sbj Prob	Over/Under	Weight
2	1	-2.16	0.095	0.600	0.505	0.025126
3	2	-2.16	0.034	0.533	0.499	0.025126
4	3	-2.16	0.024	0.533	0.509	0.025126
5	4	-2.16	0.191	0.467	0.276	0.025126
6	5	-2.16	0.272	0.667	0.395	0.025126

Figure 2. Calculating the weight statistics

Weighting the over/underconfidence scores

To apply the weighting statistic described above, simply add a column that multiplies the over/underconfidence score by the weighting score. This produces a metric of over/underconfidence that is weighted for the n size of examinees at each particular Rasch ability estimate level (θ_{ac}). By themselves, these numbers are not very informative, as they are only an intermediate step toward calculating the IDI. See Figure 3 for an example.

	A	B	C	D	E	F	G
1	Item#	Ability	Obj Prob	Sbj Prob	Over/Under	Weight	Over/Under*Weight
2	1	-2.16	0.095	0.600	0.505	0.025126	0.012688442
3	2	-2.16	0.034	0.533	0.499	0.025126	0.012546064
4	3	-2.16	0.024	0.533	0.509	0.025126	0.01279732
5	4	-2.16	0.191	0.467	0.276	0.025126	0.006926298
6	5	-2.16	0.272	0.667	0.395	0.025126	0.009916248

Figure 3. Calculating the weighted over/underconfidence scores

Finalizing the IDI calculations

To calculate the IDI for each item, one simply adds up all of the weighted over/underconfidence scores for each item. This provides a single statistic for each item that summarizes the overall amount of over/underconfidence on the item, controlling for frequency.

To quickly calculate this sum, the SUMIF function can be employed. For example, the following:

```
=SUMIF('IDI Calcs'!A:A,'Summary Stats'!A2,'IDI Calcs'!G:G)
```

instructs Excel to add up the weighted over/underconfidence scores from the IDI Calcs sheet for all rows whose item number matches that found in A2 of the Summary Stats sheet.

Interpreting the IDI

Paek et al. recommend using the differential item functioning (DIF) effect size scale recommended by Dorans and Holland (1993) to determine the size of the over/underconfidence. The scale is applied to the IDI as below:

$$\text{Large discrepancy} = IDI_i > |0.10| \tag{5}$$

$$\text{Medium discrepancy} = |0.05| < IDI_i \leq |0.10| \tag{6}$$

$$\text{Small or negligible discrepancy} = 0 < IDI_i \leq |0.05| \tag{7}$$

Therefore any IDI over 0.10 indicates a large degree of overconfidence, whereas one below -0.10 indicates a large degree of underconfidence. Medium discrepancies are found between 0.05 and 0.10, either positive or negative, and small discrepancies are those with IDIs under 0.05, either positive or negative.

Calculating the Discrepancy Percentage (DP)

Paek et al. describe two test-level summary statistics of over/underconfidence, but the first, the Test Discrepancy Index (TDI) is rather difficult to interpret, so they describe a transformation of the TDI to a simple percentage, called the DP.

The DP is expressed as follows (Paek et al., 2008, p. 5):

$$\text{Discrepancy Percentage (DP)} = \frac{\sum_{\theta_{ac}}[\sum_i P_i^* - \sum_i P_i]w(\theta_{ac})}{\text{Test Length}} \times 100 \tag{8}$$

Once again, the equation appears much more complex than the procedure actually is. In this case, we have already calculated the bulk of the numerator, as the second half of it is simply the weighted over/underconfidence scores. The rest of the numerator is simply adding all of them up for all of the items and all of the ability levels (θ_{ac}). If you have calculated them in a spreadsheet software package as described here; they are all in one column. This sum is the TDI, but we will transform it to a percentage to make it easier to interpret.

The sum of all the weighted over/underconfidence scores (aka the TDI) is divided by the test length (i.e., the number of items on the instrument), and this product is multiplied by 100 to transform it into a percentage. The result is the DP, which represents the percentage of over or underconfidence on the entire test.

Application of the method to the example dataset

Using the IDI and DP to investigate overconfidence at the item and test levels

The Paek et al. method described above was applied to the data described in the Data section. The IDIs and their effect sizes can be found in Table 1.

Table 1. ***IDIs and effect sizes for the items on the utterance verb instrument.***

Item	Word	IDI	Effect Size
1	speak	0.006	S
2	speak	0.260	L
3	speak	0.319	L
4	speak	-0.086	S
5	speak	-0.086	S
6	talk	0.255	L
7	talk	0.232	L
8	talk	0.225	L
9	talk	0.477	L
10	talk	0.318	L
11	say	0.014	S
12	say	0.013	S
13	say	0.071	M
14	say	-0.156	S
15	say	0.128	L
16	tell	0.075	M
17	tell	-0.034	S
18	tell	0.037	S
19	tell	0.001	S
20	tell	-0.017	S

The application of the Paek et al. method to these data reveals a large discrepancy between the respondents' accuracy on the items focusing on the use of the verb "talk". Since the direction of the discrepancy is positive, the IDIs are indicative of a high degree of overconfidence regarding the use of this verb. The verb "tell" seems to have little discrepancy between accuracy and confidence, and the other two verbs display a mix of over- and under confidence, as well as degree of discrepancy. The overall level of overconfidence as indicated by the Discrepancy Percentage (DP) was 10.26%, indicating that respondents were, on average, approximately 10% more confident of their answers than their actual accuracy.

Plotting the objective and subjective probabilities

Item-level plots

Another useful application of the Paek et al. method for investigating over/underconfidence on items is plotting the objective probabilities (P_i) against the subjective probabilities (P^*_i) for individual items. Doing so allows the researcher to investigate how the discrepancy changes with accuracy, and can shed a considerable amount of light upon the internal workings of examinees' minds at different levels of ability, as they encounter test items. See Figures 4 and 5 for examples.

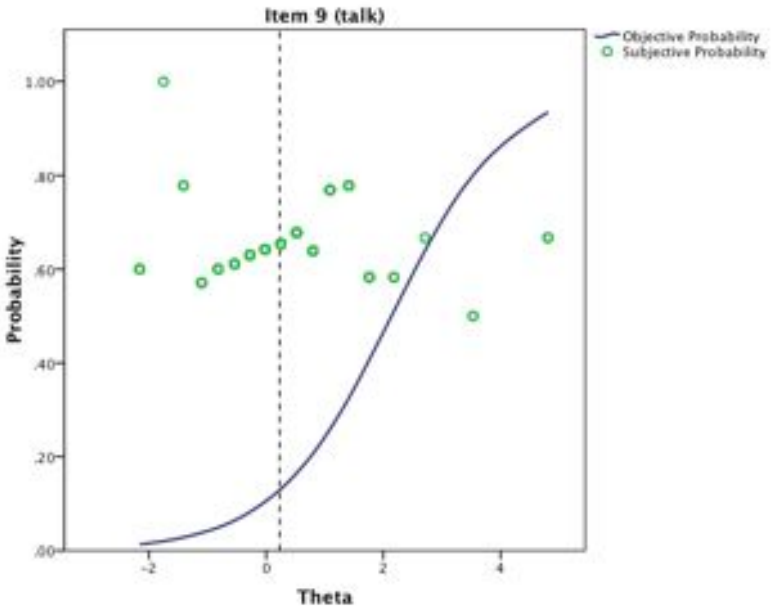

Figure 4. Plot of objective probability (accuracy) and subjective probability (confidence) against Rasch theta (ability/difficulty) for item 9. The dotted line represents the Rasch difficulty of the item.

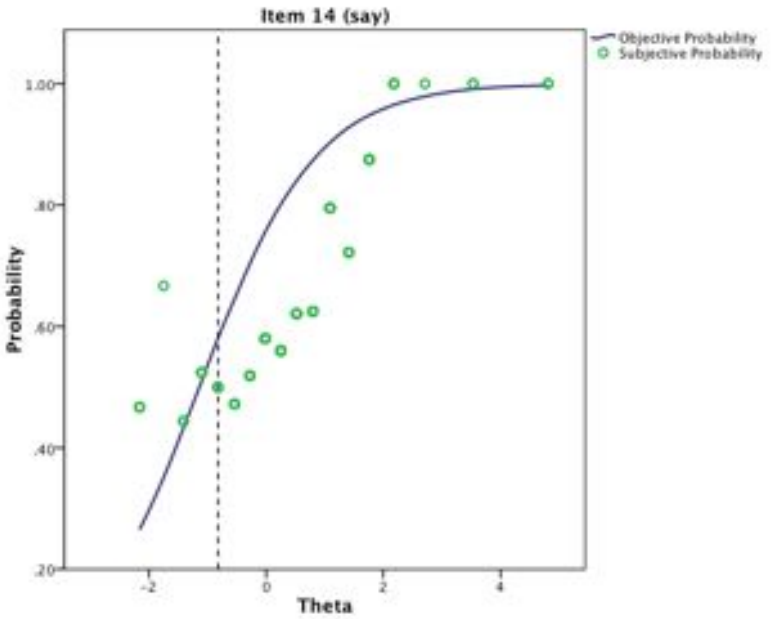

Figure 5. Plot of objective probability (accuracy) and subjective probability (confidence) against Rasch theta (ability/difficulty) for item 14. The dotted line represents the Rasch difficulty of the item.

Figure 4 is a plot of the objective probability (accuracy) and subjective probability (confidence) against the Rasch theta (the ability of the examinees and the difficulty of the items). The dotted line simply

serves as a reference point at the location of the difficulty of the item. In the case of item 9, which was the item with the highest positive IDI, indicating overconfidence, it is clear that lower-ability examinees are much more confident than their accuracy justifies. Interestingly, however, higher-ability examinees have similar confidence levels. This suggests that regardless of ability, examinees have roughly the same degree of confidence of their answers on this item. In the forthcoming paper on this study, we will discuss this further, but such a discussion is beyond the scope of this paper.

Figure 5 is a plot of item14, which was the item with the largest degree of underconfidence, as signified by the lowest IDI value. The relationship between confidence and accuracy here is interesting, in that it follows a roughly S-shaped curve, with low-ability examinees being slightly overconfident, mid-ability examinees being underconfident, and high-ability examinees' confidence and accuracy matching fairly closely.

To create the plots, a high-quality charting program is recommended. I have used the Simple Scatter/Dot chart type in SPSS 20 (IBM Corp., 2011). Both objective probability and subjective probability are assigned to the Y-axis, and the Rasch ability score is assigned to the X-axis. The objective probability plot is set to an interpolation line with the "Spline" setting, which makes interpreting the relationship between objective and subjective probabilities simpler. Finally, the optional line at the difficulty theta of the item is added manually via the "Reference Line" feature.

Other statistical or graphical packages are sure to include sufficient features to display these plots, but they are too involved for Microsoft Excel or other spreadsheet applications.

Test-level plots

Similar plots can be produced to examine the interaction between accuracy and confidence at the text level by adding up all the objective probabilities (P_i) and subjective probabilities (P^*_i) for each ability level (θ_{ac}). It is important to remind the reader that these probabilities must be from those calculated for the IDI calculations, so there is only one probability type per item, per ability level. Once again, the Excel SUMIF function is recommended here.

Once these are summed, they can be plotted against each other as in Figure 6. Once again, both the objective and subjective probabilities are assigned to the Y-axis, and the Rasch ability level (a.k.a. theta) to the X-axis.

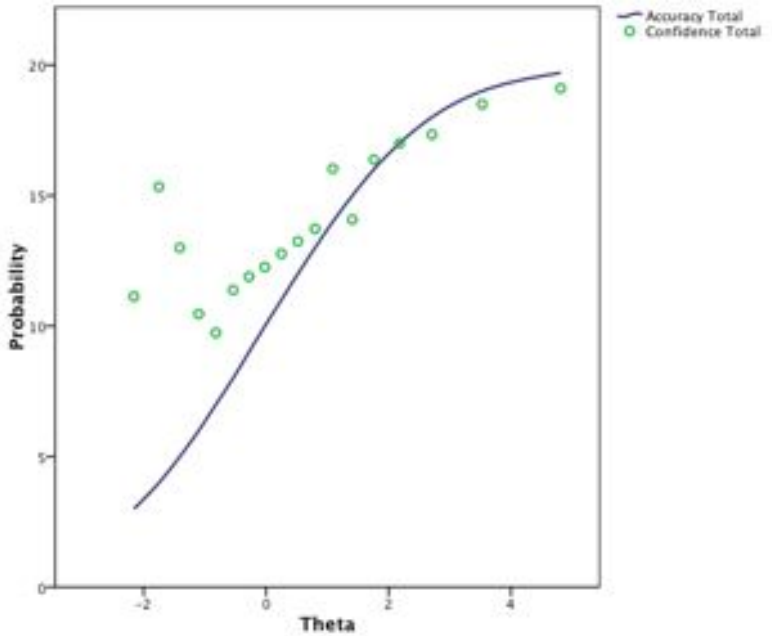

Figure 6. Plot of total objective probability (accuracy) and total subjective probability (confidence) against Rasch theta (ability/difficulty) for the entire test.

In the case of these data, it can be seen that lower-ability respondents tend to be overconfident, and higher-ability respondents tend to be underconfident. Once again, the implications of these findings within the scope of the study in question are beyond the scope of this paper, and will appear in a forthcoming paper by Sato and Batty.

Caveats regarding choice of model

Limitations of the Rasch model

One problem with the use of the Rasch model to examine confidence is that it assumes equal discrimination between items (i.e., all items discriminate between low- and high-ability examinees equally), typically with a slope of 1 (De Ayala, 2009). If one's items deviate from that assumption by a great deal, this presumption of equal discrimination can hide or otherwise mis-characterize the discrepancy between accuracy and confidence at the item level. For example, Figures 7 and 8 present the plots of probability against ability for items 9 and 14, originally discussed above, when scaled under the 2PL model instead of the Rasch model. Although 14 is similar to its Rasch counterpart, the addition of information about the discrimination of item 9 reveals the real source of the discrepancy: the item characteristic curve (ICC) is actually almost completely flat, meaning that examinees at all ability levels were roughly as likely to answer it correctly, so any interpretation of the discrepancy between accuracy and confidence here is essentially impossible. The item is simply far too hard, with a discrimination index of 0.082 and a difficulty measure of 18.905! In this case, test-level judgments based on the Rasch model may be valid, but item-level judgments would be very problematic.

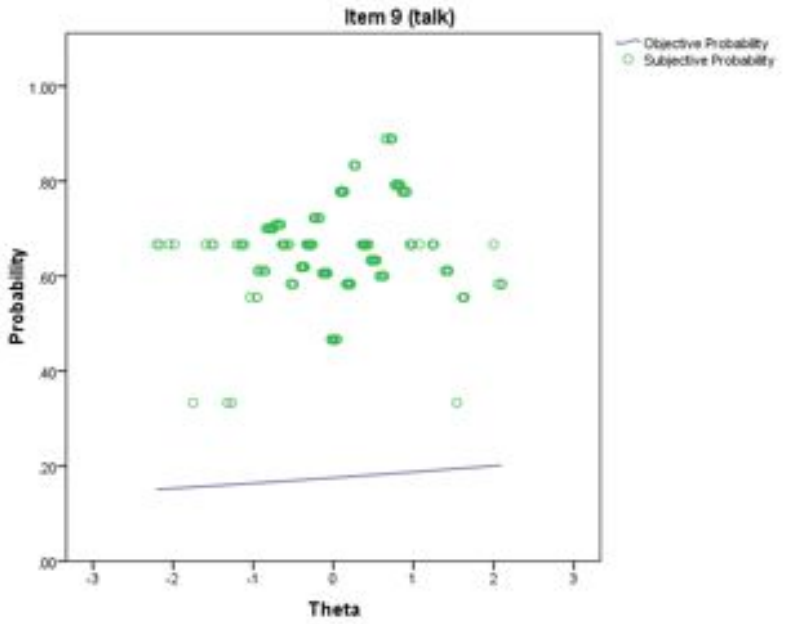

Figure 7. **Plot of objective probability (accuracy) and subjective probability (confidence) against 2PL theta (ability/difficulty) for item 9. The difficulty of the item is off the scale of the graph.**

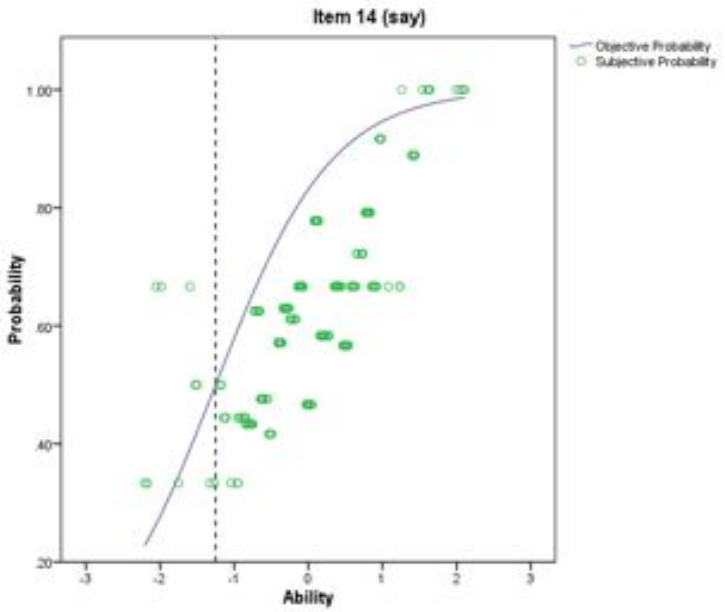

Figure 8. **Plot of objective probability (accuracy) and subjective probability (confidence) against 2PL theta (ability/difficulty) for item 14. The dotted line represents the Rasch difficulty of the item.**

Application of multi-parameter IRT models

If one concludes that a multi-parameter model would be more appropriate for the data at hand, the use of the two- or three-parameter logistic model with the Paek et al. method follows precisely the same steps as those for the Rasch model (see Stankov, Lee, & Paek, 2009). It may not be possible, however, to group respondents strictly by ability score for the calculation of subjective probability scores, as the addition of extra parameters results in fewer respondents with identical scores, rendering any kind of abstraction difficult. For the graphs presented in Figures 7 and 8, ability scores were grouped at the tenth-of-a-logit level (e.g., respondents with ability scores of 0.055 through 0.1444 would be grouped at the 0.1 level, by rounding to the nearest tenth of a logit) for the calculation of subjective probabilities and the weighting terms.

Conclusion

This paper has attempted to provide a more-accessible, practical explanation of the method developed by Paek et al. for examining the interaction between respondent confidence and accuracy using Rasch ability levels. Finally, the importance of adequately exploring one's data and considering its fit to the intended model was demonstrated.

I have focused my explanation here on only the most-straightforward of the analyses developed by Paek et al., as they are likely the most instructive and the most desired by most language-testing researchers. The original Paek et al. paper includes further analyses which require slightly more statistical knowledge to perform, but which are nonetheless very useful. Before applying the method discussed here, however, the reader is strongly encouraged to read the Paek et al. paper(s) and use the present article as a practical guide to applying the method to one's data.

Acknowledgements

This research was funded in part by Keio Gijuku Academic Development Funds (慶應義塾学事振興資金) at Keio University, Fujisawa, Japan.

References

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hakstian, A. R., & Kansup, W. (1975). A Comparison of Several Methods of Assessing Partial Knowledge in Multiple-Choice Tests: Ii. Testing Procedures*. *Journal of Educational Measurement*, *12*(4), 231–239. doi:10.1111/j.1745-3984.1975.tb01024.x

IBM Corp. (2011). *IBM SPSS Statistics*. Armonk, NY: IBM Corp.

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, *77*(3), 217–273. doi:10.1016/0001-6918(91)90036-Y

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1981). *Calibration of Probabilities: The State of the Art to 1980*.

Linacre, J. M. (2012a). Winsteps (Version 3.75.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved from http://www.winsteps.com/

Linacre, J. M. (2012b). Facets (Version 3.70.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved from http://www.winsteps.com/facets.htm

Paek, I., Lee, J., Stankov, L., & Wilson, M. (2008). *A study of confidence and accuracy using the Rasch modeling procedures* (Research Report No. RR-08-42). Princeton, NJ: Educational Testing Service.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Sato, Y., & Batty, A. (2012). A study of learners' intuitions behind the use of utterance verbs in English. *Vocabulary Learning and Instruction*, *1*(1), 29–36. doi:10.7820/vli.v01.1.sato.batty

Stankov, L., Lee, J., & Paek, I. (2009). Realism of confidence judgments. *European Journal of Psychological Assessment*, *25*(2), 123–130. doi:10.1027/1015-5759.25.2.123

Careers in Language Testing

Alistair Van Moere

Aaron Olaf Batty
abatty@sfc.keio.ac.jp
SRB Associate Editor

Alistair Van Moere is the new president of Pearson's Knowledge Technologies Group, and is responsible for the development, delivery, and validation of their automated language tests, including the Versant test (previously known as PhonePass). Prior to his employment at Pearson, Alistair was instrumental in the development of the Kanda English Proficiency Test at Kanda University of International Studies in Chiba, and drew from his experience with its speaking test in his PhD work under Charles Alderson at Lancaster University—work which won him the Jacqueline Ross TOEFL Dissertation Award in 2010.

He spoke with SRB *in the last issue about psycholinguistic assessment, and returns in this issue for our new series profiling those who have made a career of language testing. We were pleased that Alistair was once again willing to take some time out of his busy schedule to talk to us.*

About 10 years ago you were a lecturer at Kanda University of International Studies in Japan. How did you get from that position to where you are now?

Actually before Kanda University I was at Shane English Schools, first as teacher then Director of Studies. From there I moved to Kanda. Prof Frank Johnson was the Director of the English Language Institute at the time, and he created an environment where teachers could grow as researchers, and resources were made available to worthy research projects. I'm very indebted to Frank for believing in me and providing me the opportunities to succeed. I took over the coordination of the Kanda English Proficiency Test (KEPT), mainly because I was the only person, from among 35 lecturers, who was interested in doing it. Our visiting consultant, William Bonk, inducted me into language testing. I also taught myself a lot of statistics, and prepared for a PhD.

From there I did my PhD at Lancaster University with Charles Alderson, who was an excellent supervisor. I still contact him for advice. Living in Lancaster felt like exile, it's a pretty isolated place! But it's a strong department and I also learned from people like Dianne Wall, Jayanti Banerjee, and my classmate Spiros Papageorgiou. As my PhD was finishing I looked around to work in a testing company, and Ordinate Corporation in California was head and shoulders above everything else. It was a small, technology driven company, and the work going on was just so cool. It was pioneering speech processing technology and it exuded "the future". Soon after I was recruited into the management team at Ordinate we were acquired by Pearson, and so I unwittingly joined the world's largest education company, which has also been very good to me.

You won the 2010 Jacqueline Ross TOEFL Award for best PhD in language assessment. Have you got any advice for researchers undertaking PhDs in language testing?

It's all about getting the right supervisor and your relationship with your supervisor. Make sure you find someone who is responsive, who is going to read your submissions and give you feedback promptly,

and have good discussions with you. You also need to seek out and surround yourself with excellent people. If there aren't any where you are now, then move. I've been very lucky in this respect, at Kanda, Lancaster and now Pearson. For example at the moment Jared Bernstein, our Chief Scientist, is an excellent mentor. He has a wealth of experience, sees straight to the heart of problems I bring to him, and continually challenges my assumptions. Having people like this around you greatly improves the quality of your thinking.

You spoke with us in the last issue about psycholinguistic testing versus communicative testing. How do you anticipate that a testing practitioner—who is responsible for, say, running a placement test for 1,000 students in a university—can reconcile these two approaches?

Teachers have to be aware of the pros and cons of each approach. A test such as the group discussion reflects what happens in the communicative classroom, and provides washback and practice on pair- and group-work. But, it allows students to fall back on personality, practice avoidance strategies, and it's reliable for separating students into no more than two or three bands. It might also disadvantage students that haven't been inducted into that discussion format before, if they are fresh from high school. Also, score gains might be more illusory than we imagine. At Kanda, freshman students typically increased their score by a few points on a score scale of 0-20, but this was largely due to gains on the trait Communicative Effectiveness. They need only be a bit more comfortable with the interaction, and incorporate back-channeling and enthusiasm, to boost their score.

On the other hand, you can test oral proficiency with a series of discrete point items, such as sentence repeats, or reading a passage aloud and then summarizing it, and then scoring the speech for accuracy and fluency. This is a more controlled approach that creates a level playing field on which to evaluate students and allows the examiner to probe proficiency in a standardized way via items of measured difficulty. But, it's less communicative.

I'm not saying that one approach is better than the other. Just that while communicative tests appear more authentic the performances actually mix in a lot of extraneous skills, and the more reliable approach is to control and standardize the assessment. When it comes to a high-stakes assessment, we should take a mixed approach: we want the benefits of high reliability, as well as the benefits of tasks that elicit communicative skills.

You are responsible for the quality of millions of tests being taken around the world, many of which determine people's career or educational opportunities. What keeps you awake at night?

Test crackers. These are the people or agencies who make a concerted effort to take tests, memorize items, and train students to get higher test scores without improving their proficiency. We have plenty of measures in place to counter them, including biometric identity checks, monitoring during testing, and data forensics. We are also investigating whether certain item types which require a rapid, immediate spoken response are more immune to test preparation strategies. But test crackers are nevertheless a threat.

I think people tend to overlook the fact that when it comes to large-scale English proficiency testing, any test score is a combination of two abilities: language proficiency plus test preparation, and test prep is an alarmingly large proportion of the score. Pearson tests are less susceptible to this because we are a relatively new player, compared to the established test providers, and there isn't a test cracking industry built around the Versant tests or PTE Academic. Anyway, the effects of test prep and test designs which counteract test prep is an entirely under-researched area and I'd welcome more attention given to this.

What's your biggest challenge?

Finding exceptional people to hire. I am always on the lookout for test developers who are trained in linguistics and statistics, who have good project management skills, and can work in a business environment.

What do you see happening in the field these days that has you excited?

In language testing, I'm excited by any studies that involve speed or response time. I think this is an undervalued piece of information in language assessment. Two students can get the answer right, but the one who responds twice as quickly may be much more proficient than the other.

In speech processing, there is extremely exciting research in the measurement of soft skills or aspects of the speaker's state of mind. So for example the machine can predict whether the speaker is friendly, likeable, or patient. This is very promising for our clients in the customer-service (call center) industry

Any new developments since last we spoke?

We have just launched a 4-skills, certification business English test for Japan and Korea called E^Pro. It is computer-based, automatically scored, and involves interactive item-types such as responding to emails, and providing oral and written summaries. It is just 90-minutes, reports details on subskills, and has clear advantages over other certification English tests currently available. It can be taken in Pearson VUE centers (http://www.eproexam.com/).

Thank you for taking the time to speak with us!

Rasch Measurement in Language Education Part 7:

Judging plans and disjoint subsets

James Sick
International Christian University, Tokyo

Previous installments of this series have provided an overview of Rasch measurement theory, reviewed the differences among the various Rasch models, and discussed the assumptions and requirements that underlie Rasch measurement theory (RMT). In this installment, I will address a practical problem that can occur when using many-facet Rasch analysis (MFRA). MFRA is often used to adjust for differences in rater severity or other factors when measures are constructed from subjective judgments. Readers unfamiliar with MFRA and the differences among the Rasch family of models might wish to review Part 3 in this series.

Question:

My institution recently held a student speech contest with 9 teachers serving as volunteer judges. The 51 student participants were assigned to 3 rooms where a three-judge panel rated each speech for content, language, and presentation. When all speeches were completed, the scores were compiled and the three highest scoring students received a prize.

Now that the contest has finished, I am analyzing the results with MFRA with the aim of improving the judging process in future contests. When I run the analysis using Facets (Linacre, 2012a), it runs but returns the message "warning – there may be 3 disjoint subsets." Could you explain what this means and what, if anything, I should do about it?

Answer:

With some follow-up communication, we determined that the 3 judges assigned to each room did not rotate. That is, three judges rated all speeches in room 1, three different judges rated all speeches in room 2, and yet another panel of 3 judges rated the speeches in room 3. These are the 3 disjoint subsets. We can estimate the relative severity of judges within rooms by examining how they rated the same speeches. However, we cannot make similar comparisons with judges in other rooms because they rated no speeches in common.

We also confirmed that the mean scores awarded in each room differed: room 3 had a mean substantially lower than rooms 1 and 2. Was this because the speeches delivered in room 3 were of lower quality? Possibly. However, it is equally feasible that the speeches in that room were as good as the others, but the judging panel was more severe in how they interpreted and applied the judging criteria. Perhaps the three panels calibrated their scores independently at the start of the sessions. Alternatively, perhaps one judge on panel 3 had substantially more demanding standards, bringing down the average score for that room. In fact, a casual inspection of the raw scores indicated that one judge in room 3 awarded fewer points in total than any of the other 8 judges, lending support to that possibility. At any rate, because the lower scores in room 3 could feasibly be due to either judge differences or speech differences, we cannot fairly compare speech scores across rooms.

The MFRA that you conducted with Facets has used the raw scores from all performances to construct a single, logit-delineated scale, and placed both judges and participants along it. Logit measures for participants indicate the quality of their speeches. Logit measures for judges indicate their severity in applying the rating scale. The participant measures have been automatically adjusted for judge severity

by adding or subtracting an amount equal to the average severity of the judges who provided the scores. Unfortunately, in your analysis this accomplishes very little. Because all participants in a room were scored by the same three judges, the severity adjustment in any particular room will be the same for all. Moreover, differences in severity among judges in different rooms, even though Facets has estimated them, are not dependable. Facets employs a procedure called maximum likelihood estimation to locate the combination of rater and participant measures that is consistent with the data and best fits the Rasch model. However, this "best fit" solution is neither predictable nor transparent when there are disjoint subsets. The final estimate could be attributing room differences to performances, to judges, or to any additive combination of the two.

Judging Plans

Because the contest is finished and your goal is to improve judging in future speech contests, let us consider some possibilities. First of all, you could simply treat the three rooms as separate contests and award prizes to the top performers in each room. However, if the best speeches of the day happen to take place in the same room, speakers in the "strong room" would be at a disadvantage. A better approach would be to create a judging plan that rotates judges through the rooms as the contest progresses. This would link all judges, eliminate the disjoint subsets, and allow you to create a fair and dependable scale that applies to all participants independent of their room assignment.

Table 1 shows how such a judging plan would work. After 6 speeches, a short break is called and three judges rotate to other rooms. After another 6 speeches, a second set of judges rotate. With this simple plan, 6 pairs of judges would rate 11 speeches in common, 3 pairs would rate 6 speeches in common, and 3 pairs would have no common ratings but would be indirectly linked via two other judges. This would be sufficient to eliminate the disjoint subsets and create a common rating scale applicable to all rooms.

Table 1. Simple judging plan for speech contest

Session	Room 1 (Judges)	Room 2 (Judges)	Room 3 (Judges)
1 (speeches 1-6)	1 2 3	4 5 6	7 8 9
2 (speeches 7-12)	1 2 9	4 5 3	7 8 6
3 (speeches 13-17)	1 8 9	4 2 3	7 5 6

Group-anchoring

Another approach to dealing with disjoint subsets is to employ group-anchoring. Group-anchoring allows us to specify which measurement facet, in this case speakers or judges, will be considered the source of any variance between the subsets. For example, before starting the estimation process we specify that the mean speech performance measure for each room will be set to zero logits. Estimates of judge severity and individual performance within a room are then calibrated in relation to that benchmark. In effect, this forces the mean performance measures for each room to be equal, adjusting severity measures to compensate. Conversely, we could specify that the mean severity for each judging panel be set to zero. This would put faith in the judges and attribute group differences in performance measures to lower quality speeches.

Although group-anchoring may appear arbitrary, we can usually build a case that it is preferable to anchor one facet rather than the other if there are disjoint subsets. The following are some issues to consider when specifying group-anchoring:

1. *Sample Size.* Group-anchoring can take sample size into account. In the speech contest, the speakers outnumber the judges, so speech performance means are less likely to be affected by sampling error. With only three judges per room, a single strict judge, assigned by chance, can substantially skew the group mean. With 17 speakers per room, it would require about 6 weak speakers to similarly skew the mean. While 17 is hardly a robust sample, it is certainly better than 3.
2. *Incorporating additional information.* There is often anecdotal or other extraneous information to support anchoring one facet over the other. In the speech contest, one judge appeared to be quite tough based on the raw scores awarded. Apart from the speech contest, is he or she known to be a tough grader? No information was provided, but were speakers assigned to rooms randomly, or were room assignments related to classes, departments, levels, or other factors that might affect speech performances? If there are reasons to believe a priori that students in one room were of lower proficiency, one could argue for anchoring the judges rather than the speakers.
3. *Transparency.* Group-anchoring, even when wrong, creates transparency. For example, if we elect to group-anchor the speakers, we can add a caveat such as "assuming that the speeches delivered in each room were of equal quality on average, speakers 5 and 9 in room 1 and speaker 3 in room 2 delivered the best speeches of the day." Because the Rasch estimates which Facets provides are ambiguous when there are disjoint subsets, it can be advantageous to designate an hypothesized source of variance and then state the limitation. In addition, consider that the raw scores used to award prizes in the speech contest were essentially anchoring the judges. With no adjustment for judge severity, raw score comparisons assume that differences between rooms are the result of lower quality speeches. By not specifying group-anchoring, we might be accepting a default assumption that we would reject if it were made transparent.

Group-anchoring by design

In your speech contest, group-anchoring could be used, with some reservation, as a post hoc repair to compensate for a flawed design. There are instances, however, where group-anchoring can be advantageously and validly employed as part of an a priori design. To extend the discussion, let us consider the following example from an English speaking test that I helped administer several years ago.

Approximately 120 students were divided into groups of 4 and randomly assigned one of three topics for a 10-minute discussion test. Discussions were observed by two teacher-raters who did not participate in the discussion. Group assignments were quasi-random, mixing students from two or more classes, and raters were rotated frequently. The three topics were based on themes from the textbook and were known to students in advance. Topics were assigned at the start of the discussion by drawing them one by one from a canister until all had been used. This insured that the distribution of topics was equal across students and raters.

A problem with this design was that each student discussed only one topic, creating 3 disjoint subsets of unlinked topics. Consequently, it was not possible to unambiguously determine whether the topics were of equal difficulty. To exacerbate matters, there was a widespread belief among both students and teachers that Topic 3 was more challenging due to vocabulary and cognitive demands and would disadvantage students to whom it was assigned.

From a measurement perspective, the ideal solution would have been to have each student discuss two topics, preferably in different groups with members who had discussed different topics previously. This would have eliminated the disjoint subsets and allowed us to estimate the degree to which topic assignment affected performance scores. Two discussions, however, would have required an extra class period

to administer. Pedagogically, it was questionable whether the lost teaching time could be justified by minor improvements in testing accuracy.

In this context, a strong argument can be made for the validity of anchoring the students and attributing differences between subsets to topic difficulty:

1. The sample of student participants is robust. With approximately 40 randomly assigned students attempting each topic, it is reasonable to expect the mean speaking ability of each subset to closely approximate the overall mean.
2. There is an a priori prediction that Topic 3 is more difficult. If this is verified when anchoring students, it bolsters the argument that variation due to topic difficulty is the true source of any mean differences in the subset measures.

In addition, if the topic assignments are shown to affect scores, maximizing adjustments to offset this enhances face validity. Students will feel the test is more fair if they know that differences in topic difficulty are recognized and compensated for. Even though group anchoring is a compromise from a measurement perspective, the pedagogical benefits of reducing test administration time justify the cost.

Figure 1 is a Facets vertical ruler from an early administration of this test. The figure shows the relative measures of the four facets—students, judges, topics, and categories—relative to a single, logit-delineated scale along the left. As was predicted, Topic 3 is slightly more difficult than Topics 2 and 3. In comparison to the variation in rater severity and student ability, however, variations in topic and category difficulty are quite small. In practice, we found only two cases where an adjustment for topic difficulty might have been large enough to alter a student's final grade. Nevertheless, knowing that their discussion test scores took into account both the severity of the raters and the topic they were assigned seemed to increase student confidence in the overall fairness of the test. Logistically, this test would have been difficult to administer without relying on group anchoring.

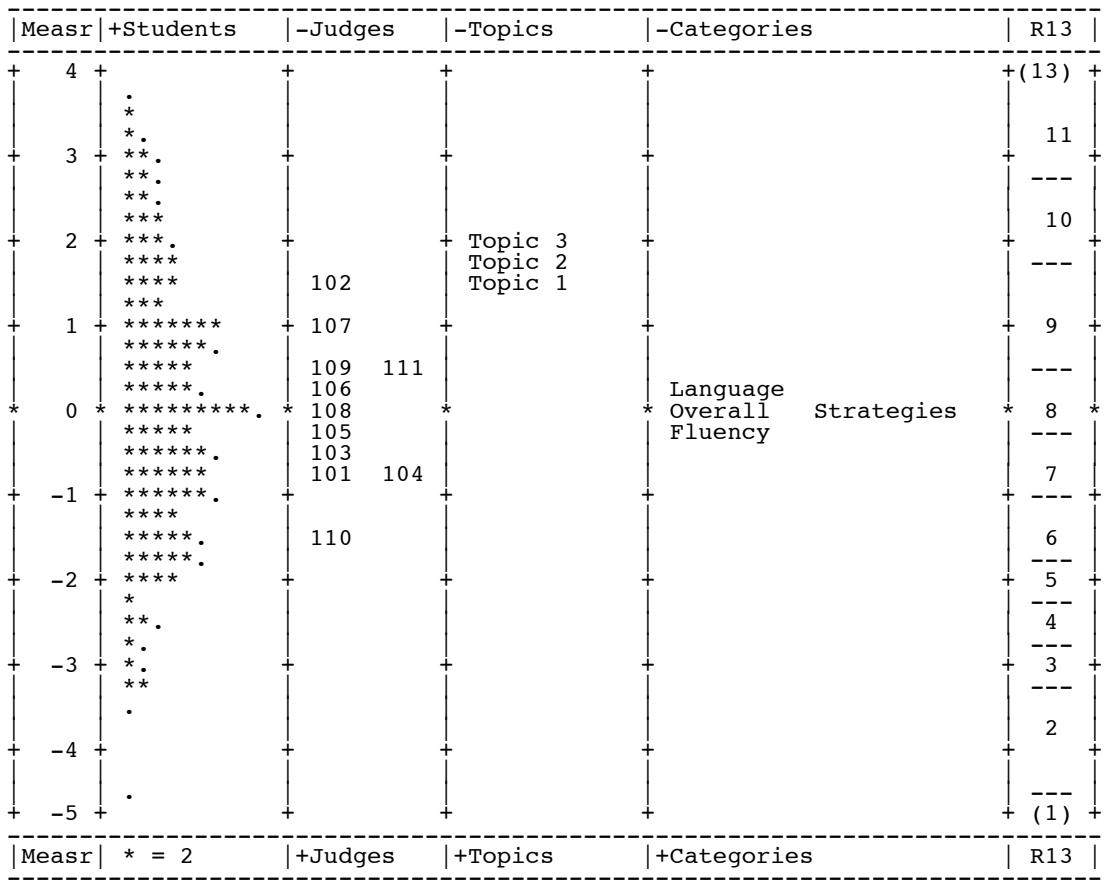

Figure 1. **Vertical ruler for a group discussion test**

More information about judging plans can be found in Linacre (1997). Details of how to specify group anchoring can be found in the Facets manual (Linacre, 2012c). An excellent tutorial on judging plans, disjoint subsets, and group-anchoring is also available at the Winsteps and Facets website (Linacre, 2012b).

References

Linacre, J. M. (1997). Judging plans and Facets. *MESA Research Note #3.*

Linacre, J. M. (2012a). Facets (Version 3.70) [Computer Software]. Chicago: Winsteps.com.

Linacre, J. M. (2012b). *Many-facet Rasch measurement: Facets tutorial #4.* Retrieved from http://www.winsteps.com/a/ftutorial4.pdf

Linacre, J. M. (2012c). *A user's guide to Facets: Program Manual 3.70.* Retrieved from http://www.winsteps.com/a/facets-manual.pdf

Statistics Corner:

Chi-square and related statistics for 2 × 2 contingency tables

James Dean Brown
University of Hawai'i at Mānoa

Question:

I used to think that there was only one type of chi-square measure, but more recently, I have become confused by the variety of chi-square measures that exist. Can you explain the difference between a simple chi-square and a (1) likelihood ratio chi-square, (2) a continuity adjusted chi-square, and (3) a Mantel-Haenszel chi-square? Finally, when should each of these statistics be used and what is the difference between a Yates and Pearson correction when used for chi-square data?

Answer:

Karl Pearson first proposed what we now call chi-square in K. Pearson (1900). Generally, chi-square (also known as Pearson's goodness of fit chi-square, chi-square test for independence, or just simply χ^2) is a test of the significance of how observed frequencies differ from the frequencies that would be expected to occur by chance, cleverly called expected frequencies. This test can be applied to many designs, but it is commonly explained in terms of how it applies to 2 × 2 contingency tables like the one shown in Exhibit 1 (below).

In order to tackle your question in more depth, I will address the following topics: calculating simple chi-square for a 2 × 2 contingency table (using an example from the literature), calculating statistics for 2 × 2 contingency tables the easy way, checking the assumptions of Pearson's chi-square, and using variations on the chi-square theme.

Calculating simple *chi-square* for a 2 × 2 contingency table

In the first study of two reported in Park, Lee, and Song (2005), the authors provide an elegant report of a 2 × 2 contingency table analysis that examined the frequency of whether apologies were present or absent in American and Korean email advertising messages. They describe their results as follows:

> Of 234 American email advertising messages, seven contained some form of apology (e.g., "We are sorry for anything that may cause you inconvenience"), whereas 74 of 177 Korean email advertising messages contained some form of apology. A chi-square test was conducted to examine the relationship between culture and the presence of apologies. The result showed that the frequency of apologies was significantly associated with culture, $\chi^2(1) = 95.95$, $p < .01$, $\varphi^2 = .23$. A greater number of Korean email advertising messages (41.81%) included apologies than did American email advertising messages (2.99%). (p. 374)

Figure 1 shows how the data need to be laid out for the calculations of the χ^2 value that Park *et al* (2005) found for the two cultures in their study (Korean and American) and the two states of Apology (present or absent).

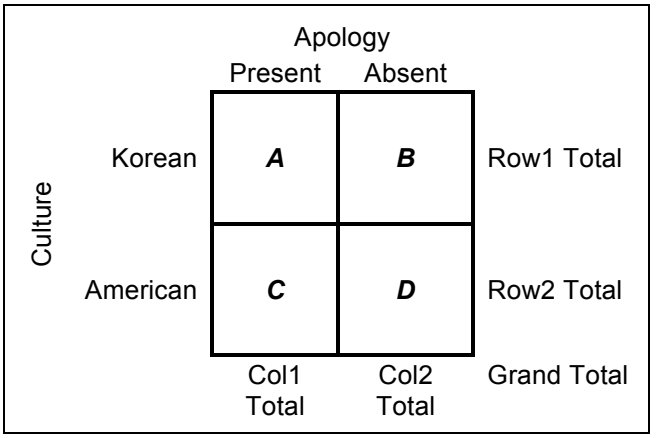

Figure 1. **Layout for Culture by Apology 2 x 2 contingency table**

In Figure 2, I have filled in the data from the Park et al (2005) study (the large numbers in italic-bold print) in the appropriate cells. Notice that the row sums on the right side in the first row are for A + B = 74 + 103 = 177 and in the second row are for C + D = 7 + 227 = 234. Similarly, the column sums at the bottom of the first column are for A + C = 74 + 7 = 81 and at the bottom of the second column are for B + D = 103 + 227 = 330. The grand total shown at the bottom right is the sum of all four cells, or A + B + C + D = 74 + 103 + 7 + 227 = 411.

Collectively, all of these values around the edges of the contingency table are known as the marginals.

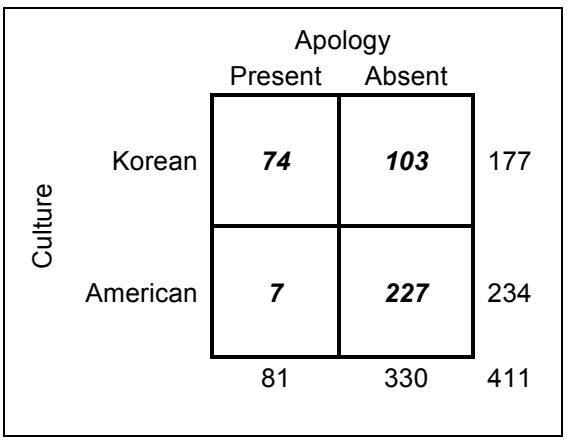

Figure 2. **Data for Culture by Apology 2 x 2 contingency table**

Observed frequencies are the frequencies that were actually found in a study and put inside the cells of the contingency table. *Expected frequencies* are estimates of the frequencies that would be found by chance in such a design (based on the marginals). Table 1 shows how the expected frequencies are calculated from the marginals for each cell. For example, the expected frequency for Cell A (Korean-Present) is calculated by multiplying the column 1 marginal times the row 1 marginal and dividing the result by the grade total, or (Col1 × Row1) / Grand Total = (81 × 177) / 411 = 34.882. The expected frequencies for cells B, C, and D are calculated similarly, as shown in Table 1. Notice that Figure 3 shows these expected frequencies in parentheses.

Table 1. ***Calculating expected frequencies***

Cell	Culture	Apology	Observed	Calculating Expected	(Col. × Row / Total) =	Expected
A	Korean	Present	74	(Col1 × Row1) / Grand Total =	(81 × 177) / 411 =	34.8832
B	Korean	Absent	103	(Col2 × Row1) / Grand Total =	(330 × 177) / 411 =	142.1168
C	American	Present	7	(Col1 × Row2) / Grand Total =	(81 × 234) / 411 =	46.1168
D	American	Absent	227	(Col2 × Row2) / Grand Total =	(330 × 234) / 411 =	187.8832

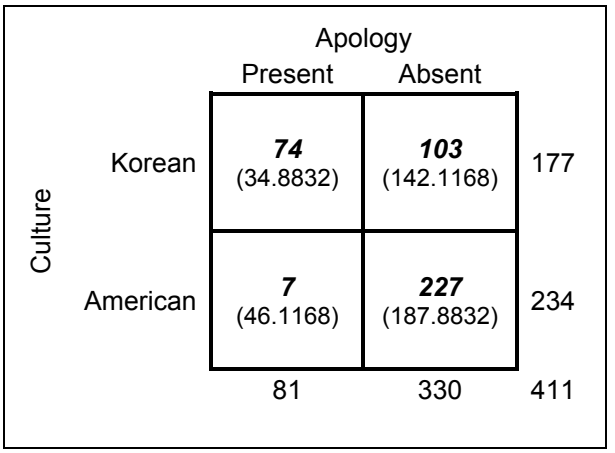

Figure 3. **Data for Culture by Apology 2 × 2 contingency table**

Table 2 shows how the *chi-square value* (χ^2) is calculated for a 2 × 2 contingency table. Intermediate values are first calculated for each cell based on the observed and expected frequencies in that cell. For example, for Cell A (Korean-Present), the value is calculated by subtracting the observed frequency minus the expected frequency and squaring the result, and then dividing the squared result by the expected frequency. In this case, that would be (Observed - Expected)^2 / Expected = (74 - 34.8832)^2 / 34.8832 = 43.8642. The same process is repeated for Cells B, C, and D as shown in Figure 3. Then the four results are summed and that sum is the chi-squared value. In this case, that would be 43.8642 + 10.7667 + 33.1793 + 8.1440 = 95.9542, or about 95.95 as reported in Park *et al* (2005).

Table 2. ***Calculating chi-square from the observed and expected frequencies***

Culture	Apology	Observed	Expected	(Observed – Expected)^2 / Expected =	
Korean	Present	74	34.8832	(74 - 34.8832)^2 / 34.8832 =	43.8642
Korean	Absent	103	142.1168	(103 - 142.1168)^2 / 142.1168 =	10.7667
American	Present	7	46.1168	(7 - 46.1168)^2 / 46.1168 =	33.1793
American	Absent	227	187.8832	(227 - 187.8832)^2 / 187.8832 =	8.1440
				Sum = ***Chi-square value*** **=**	**95.9542**

Clearly, the chi-square statistic is not difficult to calculate (though the process is a bit tedious). It is also fairly easy to interpret. As Park et al (2005) put it, "The result showed that the frequency of apologies was significantly associated with culture, $\chi^2(1) = 95.95$, $p < .01$, $\varphi^2 = .23$" (p. 374). Notice that they symbolize chi-squared as $\chi^2(1)$ [where the (1) indicates one degree of freedom] and that this chi-square value turns out to be significant at $p < .01$. [To determine the degrees of freedom and whether or not

chi-square is significant requires much more information than I can supply in this short column; however further explanations are readily available in Brown, 1988, pp. 182-194, or 2001, pp. 159-169, or at the SISA website referenced in the next section]. Note that the phi-square statistic (φ^2) that Park *et al* (2005) report will be explained below.

Calculating statistics for 2 × 2 contingency tables the easy way

Now that you understand the basic calculations and interpretation of chi-square analysis for 2 × 2 contingency tables, I will show you an easier way to calculate that statistic and all of the statistics mentioned in your question at the top of this column (as well as a few bonus statistics). The first trick is to go to the very handy *Simple Interactive Statistical Analysis* (or *SISA*) below and explore a bit:

http://www.quantitativeskills.com/sisa/

When you are ready to focus on 2 × 2 contingency table analysis go to the following URL:

http://www.quantitativeskills.com/sisa/statistics/twoby2.htm

When you arrive at that page, you will see a screen like the one shown in Figure 4. Go ahead and fill in the values from the Park *et al* (2005) 2 × 2 contingency table as shown in Figure 4. Be sure to also check the boxes next to *Show Tables:* and *Association:,* and then click on the *Calculate* button.

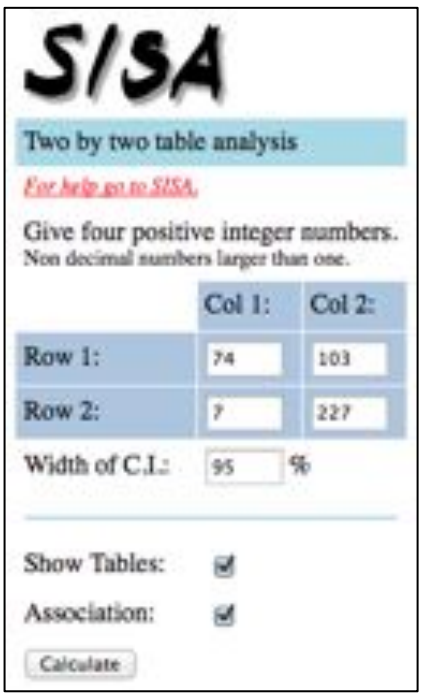

Figure 4. Using *SISA* to calculate statistics for the Park *et al* (2005) 2 × 2 contingency table

A number of tables will appear on your screen including those shown in the first column of Figure 5. You will also see numerical output like that extracted into the second column of Figure 5 (along with a good deal of additional output). Notice that the chi-squared value and its associated probability are shown in the third line of column two, labeled as *Pearson's*.

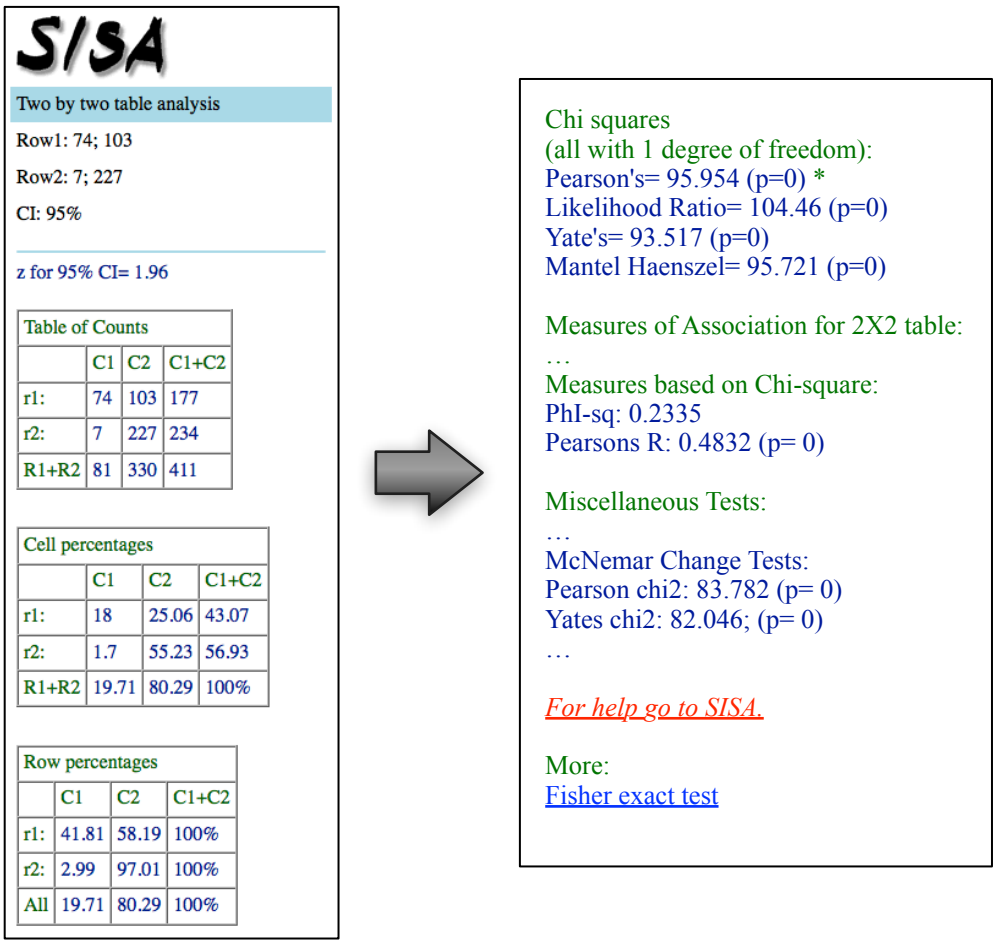

Figure 5. *SISA* output for the Park *et al* (2005) 2 × 2 contingency table

Checking the assumptions of Pearson's chi-square

Pearson's chi-square for 2 × 2 contingency tables is used to analyze raw frequencies (not percentages or proportions) for two binary variables, or put more precisely, this χ^2 statistic is a reasonable test of the significance of the difference between observed and expected raw frequencies if three assumptions are met:

1. The scales are *nominal*[1] (i.e., they are frequencies for categorical variables)
2. Each observation is *independent* of all others
3. As a rule of thumb, the *expected frequencies* are equal to or greater than 5

For example, let's consider these assumptions in the Park *et al* (2005) study. First, the scales are clearly *nominal:* culture (American or Korean) and apology (present or absent) are definitely nominal and

[1] For more on nominal, ordinal, interval, and ratio scales of measurement, see Brown (2011).

binary. Second, the observations are *independent,* which means that each observation appears in one and only one cell (i.e., each advertisement is either American or Korean and has an apology present or does not, or put another way, no advertisement appears in more than one cell). Third, the smallest *expected frequency* is 34.8832, which is well over 5. So Park et al (2005) clearly met the assumptions of Pearson's chi-square. [For more on these assumptions, see Brown, 2001, pp. 168-169.]

Using variations on the chi-square theme

Figure 5 shows selected statistics from the output that *SISA* provides. Here, I will explain the differences between these statistics, as well as when each would be appropriately applied. When the purpose of the analysis is different or the assumptions are not met, Pearson chi-square is not appropriate, but other statistics (most of which are available in the *SISA* output shown in Figure 5) have been developed for use in alternative situations as follows:

If the scales are not nominal, other non-parametric statistics (e.g., the *Mantel Haenszel Chi-square* is appropriate if both variables are ordinal—see Conover, 1999. pp. 192-194; Sprent & Smeeton, 2007, pp. 399-403; also see the SISA website) or more powerful parametric statistics may be applicable (e.g., Pearson's product-moment correlation coefficient, the *t*-test, ANOVA, regression, etc.—see Brown, 1988, 2001; Brown & Rodgers, 2002; Hatch & Lazaraton, 1991).

If the observations are not independent, Pearson's chi-square is not applicable. Period. This is a common violation that is ignored in second language research. Indeed, I searched for hours before finding the Park *et al* (2005) example that did not violate this assumption. In cases where there is a violation of this assumption, especially sequentially over time (as in a study with a dichotomous nominal variable collected from the same people on two occasions, e.g., before and after instruction), you may want to consider two other statistics: *Cochran's Q test* (see Cochran, 1950; Conover, 1999, pp. 250-258; Sprent & Smeeton, 2007, p. 215) or *McNemar's Q* (see Conover, 1999, pp. 166-170; McNemar, 1947; Sprent & Smeeton, 2007, pp. 133-135; also see *SISA* website; or to calculate this statistic: **http://vassarstats.net/propcorr.html**). In the 2 × 2 case, these Cochran's Q and McNemar's Q should lead to the same result.

If the design is larger than 2 × 2, the *likelihood ratio* (or G2) provides an alternative that can readily be used to analyze a table larger than 2 × 2 and then to examine smaller components within the table in more detail (see Sprent & Smeeton, 2007, pp. 362-363; Wickens, 1989; SISA website).

If an expected frequency is lower than five, you have three alternatives: Yates correction, the Fisher exact test, or the *N* - 1 chi-square test.

1. *Yates' correction* (Yates, 1934) is equivalent to Pearson's chi-square but with a continuity correction. In cases where an expected frequency is below 5, Yates' correction brings the result more in line with the true probability. In any case, as you can see in the second column of Figure 5, the *SISA* website will calculate this statistic for you.
2. *Fisher exact test* (Fisher, 1922) has been shown to perform accurately for 2 × 2 tables with expected frequencies below 5. The Fisher exact test (aka the Fisher-Irwin test) is more difficult to calculate than Yates' correction, but given the power of our personal computers today Yates' correction can easily be replaced by the more exact Fisher exact test. Indeed, as you can see in Figure 5, you need only click on "Fisher exact test" shown in the second column for the *SISA* website to calculate this statistic for you.
3. The *N - 1 chi-square test* is another option. Campbell (2007, p. 3661) compared chi-square analyses of 2 × 2 tables for many different sample sizes and designs and found that a statistic

suggested by Karl Pearson's son (E. S. Pearson, 1947) called the *N - 1 chi-square test* provided the best estimates. According to Campbell, as long as the expected frequency is at least 1, this adjusted chi-square (probably the "Pearson correction" referred to in the question at the top of this column) provided the most accurate estimates of Type I error levels. However, for expected frequencies below 1, he found that Fisher's exact test performed better.

If the goal is to understand the degree of relationship between two dichotomous variables, phi-square (φ^2) is calculated by dividing the Pearson chi-square value by the grand total of cases. For example, in Park *et al* (2005), $\varphi^2 = \chi^2$ / Grand total, or 95.9542 / 411 = .2335 ≈ .23. This statistic ranges from zero (if there is absolutely no association between the two variables) to 1.00 (if the association between the two variables is perfect). With reference to Figure 5, note that the phi square value is equal to the square of the Pearson correlation coefficient reported in Figure 5. In other words, squaring "Pearsons R" (.4832) in Exhibit 7 will yield a phi square of .2335.

Conclusion

As the title of this column suggested, my purpose here was to explain how chi-square and related statistics can be used for analyzing 2 × 2 contingency tables. To do so, I described the processes involved in calculating simple chi-square for a 2 × 2 contingency table (using the Park *et al*, 2005 example from the literature), calculating statistics for 2 × 2 contingency tables the easy way, checking the assumptions of Pearson's chi-square, and using variations on the chi-square theme. Along the way, I believe I addressed all parts of the original question at the top of the column.

All in all, in my experience, this family of statistics has been much abused and misused in our field—perhaps more than any other. Please consider such analyses very carefully when using them and apply them correctly. Be sure, for example, to review the correct procedures as described in some of the references listed below.

References

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

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

Brown, J. D. (2011). Statistics Corner: Questions and answers about language testing statistics: Likert items and scales of measurement? *SHIKEN: JALT Testing & Evaluation SIG Newsletter, 15*(1)*,* 10-14. Also available on the Internet at http://jalt.org/test/PDF/Brown34.pdf .

Brown, J. D., & Rodgers, T. (2002). *Doing second language research.* Oxford: Oxford University Press.

Campbell, I. (2007). Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. *Statistics in Medicine, 26,* 3661-3675.

Cochran W. G. (1950). The comparison of percentages in matched samples. *Biometrika, 37,* 256-266.

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York: Wiley.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society, 85*(1), 87-94.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics.* Rowley, MA: Newbury House.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153-157.

Park, H. S., Lee, H. E., & Song, J. A. (2005). "I am sorry to send you SPAM": Cross-cultural differences in use of apologies in email advertising in Korea and the U.S. *Human Communication Research, 31*(3), 365-398.

Pearson, E. S. (1947). The choice of statistical tests illustrated on the interpretation of data classed in a 2 x 2 table. *Biometrika, 34,* 139-167.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, Series 5, 50*(302), 157-175.

Sprent, P., & Smeeton, N. C. (2007). *Applied nonparametric statistical methods* (4th ed.). Boca Raton, FL: Chapman & Hall.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Yates F. (1934). Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society Supplement, 1,* 217-235.

Where to Submit Questions:

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

Your question can remain anonymous if you so desire.

Upcoming Language Testing Events

The 35th Language Testing Research Colloquium (LTRC): July 3 – 5, 2013

Abstract submission deadline: (closed)

Venue: Seoul National University, Seoul, Korea

Conference homepage: http://www.ltrc2013.or.kr/

2013 Pacific Rim Objective Measurement Symposium (PROMS): August 1 – 6, 2013

Abstract submission deadline: (closed)

Venue: National Sun Yat-sen University, Kaohsiung, Taiwan

Conference homepage: http://www.education.nsysu.edu.tw/TERA-PROMS2013/

The 17th Japan Language Testing Association (JLTA) National Conference: September 21, 2013

Abstract submission deadline: June 9th, 2013 (Extended)

Venue: Waseda University, Shinjuku, Tokyo

Conference homepage: https://e-learning-service.net/jlta.ac/

The 15th Midwest Association of Language Testers (MwALT) Conference: September 21, 2013

Abstract submission deadline: May 31st, 2013

Venue: Michigan State University, East Lansing, MI, USA

Conference homepage: http://sls.msu.edu/mwalt2013/about/

The 12th East Coast Organization of Language Testers (ECOLT) Conference: April 10 – 11, 2014

Abstract submission deadline: September 30th, 2013

Venue: Georgetown University, Washington, DC, USA

Conference homepage: http://www.cal.org/ecolt/

The 5th Association of Language Testers in Europe (ALTE) International Conference: October 25, 2013

Abstract submission deadline: June 3rd, 2013

Venue: Maison Internationale, Cité Internationale Universitaire de Paris, Paris, France

Conference homepage: http://events.cambridgeenglish.org/alte-2014/index.html

Shiken Research Bulletin Editorial Board

Submissions

If you have a paper that you would like to publish in *Shiken Research Bulletin,* please email it in Microsoft Word format to the General Editor at:

jaltteval+srb@gmail.com