jalt journal

The research journal of the Japan Association for Language Teaching

Volume 34 • No. 1 • May 2012

JALT

全国語学教育学会

Japan Association for Language Teaching

¥950 ISSN 0287-2420

JALT Journal

Volume 34 • No. 1
May 2012

Contents

May 2012
Volume 34 • No. 1

Japan Association for Language Teaching

A Nonprofit Organization

The Japan Association for Language Teaching (JALT) is a nonprofit professional organization dedicated to the improvement of language teaching and learning in Japan. It provides a forum for the exchange of new ideas and techniques and a means of keeping informed about developments in the rapidly changing field of second and foreign language education. Established in 1976, JALT serves an international membership of approximately 3,000 language teachers. There are 35 JALT chapters, all in Japan, along with 22 special interest groups (SIGs) and four forming SIGs. JALT is one of the founders of PAC (Pan-Asian Consortium), which is an association of language teacher organizations in Pacific Asia. PAC holds regional conferences and exchanges information among its member organizations. JALT is the Japan affiliate of International TESOL (Teachers of English to Speakers of Other Languages) and is a branch of IATEFL (International Association of Teachers of English as a Foreign Language).

JALT publishes *JALT Journal,* a semiannual research journal; *The Language Teacher*, a bimonthly periodical containing articles, teaching activities, reviews, and announcements about professional concerns; and the annual *JALT International Conference Proceedings*.

The JALT International Conference on Language Teaching and Learning and Educational Materials Exposition attracts some 2,000 participants annually and offers over 600 papers, workshops, colloquia, and poster sessions. Each JALT chapter holds local meetings and JALT's SIGs provide information and newsletters on specific areas of interest. JALT also sponsors special events such as workshops and conferences on specific themes, and awards annual grants for research projects related to language teaching and learning. Membership is open to those interested in language education and includes automatic assignment to the nearest chapter or the chapter you prefer to join, copies of JALT publications, and reduced admission to JALT-sponsored events. JALT members can join as many SIGs as they wish for an annual fee of ¥1,500 per SIG. For information, contact the JALT Central Office or visit the JALT website at <www.jalt.org>.

JALT National Officers, 2011–2012

President: . . . . . Kevin Cleary
Vice President: . . . . . Nathan Furuya
Auditor: . . . . . Caroline Lloyd
Director of Treasury: . . . . . Oana Cusen
Director of Records: . . . . . Aleda Krause
Director of Program: . . . . . Steve Cornwell
Director of Membership: . . . . . Buzz Green
Director of Public Relations: . . . . . Michael Stout

Chapters

Akita, East Shikoku, Fukui, Fukuoka, Gifu, Gunma, Hamamatsu, Himeji, Hiroshima, Hokkaido, Ibaraki, Iwate, Kagoshima, Kitakyushu, Kobe, Kyoto, Matsuyama, Miyazaki, Nagasaki, Nagoya, Nara, Niigata, Oita, Okayama, Okinawa, Omiya, Osaka, Sendai, Shinshu, Shizuoka, Tokyo, Toyohashi, West Tokyo, Yamagata, Yokohama.

Special Interest Groups

Bilingualism; Business English; College and University Educators; Computer-Assisted Language Learning; Critical Thinking (forming); Extensive Reading; Framework and Language Portfolio; Gender Awareness in Language Education; Global Issues in Language Education; Japanese as a Second Language; Junior and Senior High School Teaching; Learner Development; Lifelong Language Learning; Literature in Language Teaching (forming); Materials Writers; Other Language Educators; Pragmatics; Professionalism, Administration, and Leadership in Education; Speech, Drama, & Debate (forming); Study Abroad; Task-Based Learning; Teacher Education and Development; Teachers Helping Teachers; Teaching Children; Testing and Evaluation; Vocabulary (forming).

JALT Central Office

Urban Edge Building, 5F 1-37-9 Taito, Taito-ku, Tokyo 110-0016, Japan
Tel.: 03-3837-1630; Fax: 03-3837-1631; Email: jco@jalt.org;
Website: <www.jalt.org>

In this Issue

Articles

This issue of *JALT Journal* features a diverse range of articles. The first is a critical review of perceptual learning styles constructs and their associated instrumentation in applied linguistics by **Ian Isemonger**. This article has implications not only for present-day practitioners and researchers in our field, but for future research trajectories and journal editors as well. Our second feature article, a study by **Rie Koizumi** and **Akiyo Hirai**, compares the Story Retelling Speaking Test (SRST) with two commercial speaking tests to examine the validity of score-based interpretation of the SRST. The third article, a comparative study of personal pronoun use in argumentative essays by **Sayo Natsukari**, investigates the use the first person singular *I* by Japanese EFL learners and native English speakers. The fourth article, by **Farahman Farrokhi**, **Rajab Esfandiari**, and **Edward Schaefer**, is a study with implications for the use of peer- and self-rating in L2 writing assessment. The researchers use Many-Facet Rasch Measurement (MFRM) to investigate rater severity/leniency among three rater types—self-assessor, peer-assessor, and teacher assessor.

Reviews

Six book reviews are published in this issue. In the first, **James Essex** reviews an edited volume on the spread of English throughout Asia. Our second review, by **William Green**, is an examination of a collection of contributions from the proceedings of a conference on the teaching of English to young learners. The third review, by **Rob Higgins**, considers a practitioner-oriented EAP resource guide. Next, **Jung In Kim** reviews a volume dedicated to early language learning and teaching in a wide range of ESL and EFL contexts. The fifth review, by **Tim Murphey**, examines an introductory volume on sociocultural theory. The final review comes from **Christopher Starling** and **Yumi Tanaka**, who report on a work exploring Japanese learners' attitudes to English varieties.

From the Editor

This issue of *JALT Journal* comes with significant changes to our Editorial Advisory Board (EAB). First of all, I am pleased to welcome **Melodie Cook** as the new Associate Editor. Melodie has been working with me on screening decisions since last fall and has taken on an increasingly prominent role on

editorial tasks overall. *JALT Journal* had been without an Associate Editor for several issues and having Melodie on board means that the *Journal* will be in good hands beyond 2012. **Greg Sholdt** is also taking on a greater role on the EAB. As Consulting Editor, he will be able to more regularly contribute his quantitative expertise. Joining our EAB as well are **Bill Acton**, **Yuko Goto Butler**, **Neil Cowie**, **Tim Greer**, and **Tamara Swenson**. Bill, of Trinity Western University in Canada, is a former JALT Publications Board Chair. Yuko was a plenary speaker at the JALT2009 Conference and specializes in young learner FL education. She teaches at the University of Pennsylvania. Neil has been one of our most active additional readers over the past few years. Based at Okayama University, he specializes in motivation and identity. Tim has served the *Journal* well in recent years as an additional reader and his expertise in Conversational Analysis is most welcome. He teaches at Kobe University. Tamara, of Osaka Jogakuin College, is a former *JALT Journal* Editor and therefore brings valuable editorial experience to the board in addition to being able to review papers related to content-based instruction and other areas. We extend a warm welcome to each of you. Thanks go out, as always, to the ongoing members of the Board, especially those who have contributed to this issue.

JALT Journal is deeply indebted to four members of the Board who are leaving after a long period of service. **Bill Bradley**, **Eli Hinkel**, and **Mary Gobel Noguchi** have been reviewing papers and providing editorial counsel since the 1990s, and I cannot say enough about their contributions. Special thanks as well to **Nick Jungheim** who is also leaving the Board. A former Editor of *JALT Journal*, Nick will be especially missed. The editors who have followed him have relied on his leadership and support over the years. We hope and trust that each of the outgoing EAB members will continue to support *JALT Journal* and the JALT organization.

There are cosmetic changes to *JALT Journal* as well. Our old cover style, which we have used since the May 2000 issue, is being replaced with this new solid British racing green. Rather than changing the color each year, we have opted for the permanent, one-color cover most commonly used by research journals, moving the contents information inside and adding the JALT logo. We hope you like this new look for *JALT Journal*. Thanks to **Malcolm Swanson** for his creative insights in designing this new cover.

I wish to express my sincere thanks to **Aleda Krause** and the members of the *Journal* Production team—**Amy Brown, Susan Gilfert, Joseph Sheehan, Alan Stoke**, and **Jack Yohay**—for their outstanding work. Susan is the newest member of the team, and I would like to officially welcome her with this issue.

Darren Lingley

Articles

Perceptual Learning Styles and Lessons in Psychometric Weakness

Ian Isemonger
Kumamoto University

In this critical review, I argue that the usefulness of perceptual learning styles constructs within applied linguistics is very limited. Researchers within applied linguistics have neglected to engage with objections to these constructs which date back to the 1970s within general educational research. Problems of poor instrumentation are considerable and predictive power has not sufficiently been demonstrated. It is argued that these constructs present a special case for measurement because they are not easily operationalized through statements on self-report questionnaires. I discuss implications for practitioners and research, and argue for greater editorial oversight in preventing poor instruments from entering the literature in the future. Some specific recommendations which may assist with such prevention are discussed. These include a more critical approach to the use of Cronbach's alpha, the use of Confirmatory Factor Analysis (CFA) as one powerful tool to demonstrate unidimensionality, and the avoidance of paraphrased items.

本論文は、応用言語学で用いられる知覚学習スタイルの構成概念が有用性に欠けることを指摘する。一般的な教育学研究の分野で1970年代から批判されてきたこれら構成概念について、応用言語学研究者たちはその対応法の検討を十分には行なっていない。知覚学習スタイルの測定方法には問題が多く、その妥当性も明確になっていない。さらに、アンケートのような自己報告を用いた測定方法では、このような構成概念は適切に定義づけることが難しいことも指摘されている。外国語教育実践と応用言語学研究のためにも、不十分な測定方法が将来的に研究に入り込まないようにする必要があり、そのためには学術誌等の編集者によるさらなるチェック体制の強化が不可欠である。具体的方法として、クロンバック α 係数の使用に関する注意喚起や、1 次元性を確認するのに有用な確証的因子分析(CFA)の活用を提案する。

JALT Journal, Vol. 34, No. 1, May 2012

Perceptual learning styles emerged as a significant branch of study within applied linguistics during the 1980s. This was led by a groundbreaking paper (Reid, 1987) using the Perceptual Learning Styles Preference Questionnaire (PLSPQ). There was a surge of interest in the area with a number of studies employing the PLSPQ (Frank & Hughes, 2002; Hyland, 1993; Isemonger & Sheppard, 2003; Kelly, 1998; Kinsella, 1995b; Melton, 1990; O'Donoghue, Oyabu, & Akiyoshi, 2001; Oxford, 1995; Oxford & Anderson, 1995; Peacock, 2001; Shen, 2010; Siew Luan & Ngoh, 2006; Stebbins, 1995; Thomas, Cox, & Kojima, 2000; Yamashita, 1995; Yu-rong, 2007). This interest was arguably driven by the intuitive or common-sense appeal of perceptual learning styles as a potential area of individual and cross-cultural differences. Also, shortly after Reid's release of the PLSPQ, three further perceptual learning styles instruments emerged. The first of these, the Learning Channel Preference Checklist (LCPC; O'Brien, 1990) was revised by the same author 12 years later (O'Brien, 2002). The second, the Style Analysis Survey (SAS; Oxford 1993a; 1993b) measures other aspects of learning styles in addition to preferences for perceptual learning styles. The third, the Perceptual Learning Preference Survey (PLPS; Kinsella, 1995a) has far less exposure in the literature. In addition to these earlier instruments, DeCapua and Wintergerst (2005) and Wintergerst, DeCapua, and Verna (2002; 2003) have more recently sought to revise the PLSPQ in the form of the Learning Styles Indicator (LSI).

Unfortunately, the initial appeal of perceptual learning styles research has not been matched by tangible gains for applied linguistics. This disappointment is the result of a general omission to engage with the predictive power of these constructs in sound empirical studies and, above all, to address the psychometrics of scores generated by scales purporting to measure the constructs themselves—an issue which methodologically precedes the issue of predictive power. Twenty-five years on from the publication of Reid's (1987) paper, my purpose in this critical review is to scrutinize the emergence and development of the ensuing research trajectory and its associated line of instrumentation. A case is presented for where it went wrong and what can be learned from some of the mistakes. Such an assessment is important for future areas of research growth within applied linguistics and is also important in alerting both researchers and practitioners to the apparent limitations of these constructs.

The paper begins by drawing the reader's attention to the remarkable manner in which the weaknesses of perceptual learning styles research within applied linguistics parallel weaknesses and controversy within gen-

eral educational research occurring about a decade earlier. This brief detour into an instructive past, neglected by perceptual preference researchers within applied linguistics up to now,[1] is offered to strengthen the arguments for a change of thinking on the usefulness of these constructs for both research and classroom practice. Following this, the emergence of perceptual learning styles instruments within applied linguistics is considered from the perspective of psychometric credibility and predictive power. In a more positive mode of critique, suggestions for editorial oversight that might help prevent such weakness in the future are offered, and the paper then turns to some specific issues and associated guidelines for promoting more empirically secure instrumentation and its inclusion in the literature in the future. These suggestions, while important, are not intended to be exhaustive or a comprehensive treatment of the issues. Finally, attention is drawn to the above-mentioned attempts by DeCapua and Wintergerst (2005) and Wintergerst et al. (2002; 2003) to revise the PLSPQ in the form of the LSI. It is argued that their approach risks compounding the problems by retaining some of the inherent flaws of the PLSPQ and thereby perpetuating the life of an instrument, and indeed a line of instrumentation, that should be consigned to applied linguistics' psychometric past.

Preferences for Perceptual Modalities and Repeating History

The initial interest in perceptual learning styles within applied linguistics almost 25 years ago was at least partly informed by theory and constructs already in use concerning preferences for different modalities of perception. These constructs included the visual, auditory, and tactile modalities of perception and were initially operationalized through self-report by R. Dunn and K. Dunn in the 1970s in work that has covered three or four decades. This work is represented in the following contemporary instruments: the Learning Styles Inventory (Price & Dunn, 1997) or LSI (not to be confused with the Learning Styles Indicator cited above) and the Productivity Environmental Preference Survey (PEPS; Price, Dunn, & Dunn, 1996). The LSI and PEPS are designed to measure more than perceptual constructs, but the perceptual constructs contained in them prompted Reid's research (Dunn, 1983, 1984; Dunn & Dunn, 1972, 1979; Dunn, Dunn, & Price, 1975, 1978, 1979; Price, Dunn, & Sanders, 1980; as cited in Reid, 1987). The dissemination of a new line of instrumentation in the form of the PLSPQ, the SAS, the LCPC, and the PLPS within applied linguistics, purporting to measure the same perceptual modalities as those measured in the earlier work of the Dunns, seems to have occurred despite substantial objections to the viabil-

ity of this measurement, and the predictive power of what is supposedly measured, in the late 1970s and early 1980s.

Kampwirth and Bates (1980) and Tarver and Dawson (1978), in secondary research examining a number of studies, considered interaction effects between modality preference and teaching strategies and found no convincing empirical support for such interactions. Deverensky (1978) was the first to propose that the problem might be operational, arguing that the task of finding sensitive measures of such preferences was difficult. In a sequence of rebuttal and rejoinder a decade later which involved Kavale and Forness (1987; 1990) and R. Dunn (1990), the predictive power of modality preference was once more brought into question. Kavale and Forness (1987) performed a meta-analysis of 39 studies and concluded that the effect size of the interventions was small. Critically for assertions that will later be made in this paper, they also argued that the measurement of modality preference was difficult. This was consistent with observations made by Deverensky 10 years previously.

The point of recounting the above debate[2] is to demonstrate that even as perceptual learning styles research got off the ground in applied linguistics in the late 1980s, there was an existing rebuttal to answer to concerning the operational viability of such constructs and their predictive power. My own survey of the literature within applied linguistics indicates that, to the best of my knowledge, at no point has any perceptual learning styles researcher engaged directly with these early objections from outside the field. It seems that initiators of perceptual learning styles research within applied linguistics either did not know about the debate or neglected to engage with it. For whatever reason, the omission was significant because the contours of perceptual learning styles research within applied linguistics were fated to retrace those from outside of the field in a notable case of repeating history. A new set of constructs would be offered to the applied linguistics community along with companion instrumentation. No convincing evidence of predictive power would then be demonstrated for these constructs. And later, the operational viability of these constructs would be drawn into question.

Perceptual Learning Styles Research in Applied Linguistics and Psychometric Weakness

Reid (1987), as stated above, is widely considered a seminal paper, if not the seminal paper, in perceptual learning styles research within applied linguistics in that it introduced the PLSPQ to applied linguistics literature,

and its presence ever since has been significant in this literature (Bowman, 1996; Hyland, 1993; Isemonger & Sheppard, 2003; Kim, 2001; Melton, 1990; O'Donoghue et al., 2001; Peacock, 2001; Reid, 1998a, 2000; Rossi-Le, 1995; Siew Luan & Ngoh, 2006; Stebbins, 1995; Yu-rong, 2007). A search on the Internet will also reveal frequent use in unpublished postgraduate theses, symposia, and other forums. This does not include use in action research and classroom practice, the prevalence of which is difficult to determine empirically.

The PLSPQ is claimed to measure six constructs relating to learning preference, four of which are perceptual (Visual, Auditory, Kinesthetic, and Tactile) and two of which are social (Group and Individual). Reid used the instrument in 1987 without validation of scores. A subsequent (1990) paper by Reid, in the same journal, dealt with validation by reporting Cronbach's alphas, but these were not fully reported for all constructs on the final version of the instrument seen in the 1987 study and still in use today. The 1990 paper also chronicled a problematic development process for the instrument involving construct-related difficulties that have never been overcome.

The pervasive and continuing use of the PLSPQ presents a case study in the cautioning issued by Wilkinson and the American Psychological Association (APA) Task Force on Statistical Inference (1999, p. 596) concerning the tendency for defective measures to remain in use once they have entered the literature. Reid's (1990) article provided an open and forthcoming account of the problematic development of the instrument and the shortcuts taken that should have arrested its further use pending revision and demonstration of the capacity to generate valid scores. Unfortunately, the instrument had already gained momentum in the literature and its use persisted and even grew. The most significant challenge to the capacity of the instrument to produce psychometrically valid scores came as late as 14 years later (Wintergerst, DeCapua, & Itzen, 2001) in a study that employed Exploratory Factor Analysis (EFA) and Cronbach's alphas as diagnostics—although Itzen (1995) had examined the issue earlier in a dissertation that gained little exposure and that employed the additional method of Confirmatory Factor Analysis (CFA). The Cronbach's alphas from all of these studies are available for inspection in Table 1. Also included is the more recent study by Isemonger and Sheppard (2007) which employed both the methods of EFA and CFA in addition to reporting alphas.

Table 1. Comparative Alphas for Past Studies of Scores Generated by the PLSPQ

Authors	Language	Sample	V	A	K	T	G	I
Itzen (1995)	English	92 NSs	.47	.46	.66	.76	.88	.78
		126 NNSs	.54	.56	.63	.72	.87	.80
Wintergerst et al. (2001)	English	100	.37	.39	.69	.59	.87	.75
Isemonger and Sheppard (2007)	Korean	691	.37	.39	.76	.67	.83	.84
Strength			**Weak**		**Moderate**		**Strong**	

NS = Native Speaker; NNS = Nonnative Speaker

V = Visual, A = Auditory, K = Kinesthetic, T = Tactile, G = Group, I = Individual

The following observations are pertinent. The part of the PLSPQ that measures perceptual modality performs marginally in the case of the Kinesthetic and Tactile scales and poorly for the Visual and Auditory scales. Assuming Nunnally and Bernstein's (1994) criterion of .7 for scale reliability,[3] the Visual and Auditory scales' alphas are inadequate. In terms of this same criterion, the Kinesthetic and Tactile scales are marginal in performance. Furthermore, the studies conducted by Wintergerst et al. (2001) and Isemonger and Sheppard (2007) present results which threaten the claim for the discriminant validity of the Kinesthetic and Tactile scales. EFAs conducted by Isemonger and Sheppard and Wintergerst et al. failed to reduce to simple structure in line with the scoring model offered for the instrument. Finally, CFAs[4] conducted by Isemonger and Sheppard and Itzen (1995) failed to confirm Reid's six-scale model.

Turning from the PLSPQ to two of the other perceptual leaning style instruments, the LCPC (Learning Channel Preference Checklist; O'Brien, 1990, 2002) and the SAS (Style Analysis Survey; Oxford, 1993a, 1993b), the omission to demonstrate valid scores in sound psychometric studies was a concomitant feature of the emergence of these instruments in the literature. In view of the problematic psychometrics of scores generated by the PLSPQ in studies cited above, Isemonger and Watanabe (2007) conducted research into the psychometrics of scores generated by the perceptual component of the SAS. The stated goal of the research was to assess whether operational problems were specific to the PLSPQ or a possible feature of the general line of instrumentation. Results for scores on the SAS were poor. Values for

Cronbach's alpha were as follows: Visual, .69; Auditory, .56; and Hands-On, .58. Given that these are 10-item scales[5] one would expect the alpha values to be considerably higher. Furthermore, in the same study, the model hypothesized by Oxford's design and scoring regime for the instrument was not confirmed in a CFA, and an EFA indicated problems with operationalizations including labels for constructs that were wider than the operational bandwidth of the construct. This study extended the problematic from the PLSPQ to other instruments measuring such constructs. Another by Isemonger (2008) examining the psychometrics of scores generated by the LCPC has brought these instruments further into question. Cronbach's alphas in this study were as follows: Visual, .52; Auditory, .42; and Haptic, .51. Again, given that these are 12-item scales, one would expect the value for alpha to be considerably higher. The model offered in the scoring regime for the instrument was not confirmed in a CFA. I am not aware of any prior research that had comprehensively examined the psychometrics of a set of scores generated by these instruments, using either EFA or CFA as the method.

While not as widely used as the PLSPQ, the LCPC and the SAS have been employed in both research and applied practice. The LCPC is commercially available (Specific Diagnostics Inc.) and its emergence in applied linguistics literature came through Reid (1995) in a regularly cited book directed at the practitioner. It has appeared in research (Hughes, 2001; Oxford, Young, Ito, & Sumrall, 1993). The first version received a negative review in the BUROS Institute of Mental Measurements' publication (Deaton, 1992). My own research (Isemonger, 2008) used the modified 2002 version. The SAS entered applied linguistics literature through Reid's (1995; 1998b) books—the 1998 book also being directed at the practitioner. It is also available for students to self-administer on the University of Alabama's College of Arts and Science's website (Oxford, 1993b). The instrument has entered the literature pertaining to Japanese as a foreign language (Ehara, 1998) and has also found use in recent research (Henry-Vega, 2004; Psaltou-Joycey & Kantaridou, 2011). Again, use in action research and classroom practice is hard to assess empirically but, given entrance into the literature, such usage is presumed. With regard to perceptual learning styles, therefore, the current situation within applied linguistics is one of considerable concern. Instruments have entered the literature and been used over a period of 20 to 25 years without sufficient documentation of the development process, and have even been launched without provision of minimal indexes of reliability such as Cronbach's alpha. None of the instruments (PLSPQ, SAS, nor LCPC) entered the literature accompanied by results from an EFA or CFA

to justify the model (implicitly hypothesized in the scoring regime for each respective instrument) at launch time.[6] Furthermore, the limited research that has been done by independent researchers, after the introduction of the instruments, provides no reassurance. In fact, such research exacerbates the doubt.

Perceptual Learning Styles and Evidence of Predictive Power

As explained above, one of the central objections to perceptual learning styles constructs prior to their emergence as an area of interest within applied linguistics was that they lacked predictive power (Deverensky, 1978; Dunn, 1990; Kampwirth & Bates, 1980; Kavale & Forness, 1987, 1990; Tarver & Dawson, 1978). Given the strength of the objections, it would have been appropriate for applied linguistics to engage with this issue directly and from the outset—of course, after establishing good psychometrics which is necessarily prior. However, what we have seen is a research trajectory that is almost completely descriptive in nature. Very few studies have attempted to demonstrate the predictive power of these constructs in terms of learning outcome, and those that have introduced an achievement criterion remain correlational studies and, anyway, have pointed to little correlation. For example, Ehrman and Oxford (1995) found no significant correlations between learning style constructs measured by the Learning Styles Profile (LSP; Keefe, Monk, Letteri, Languis, & Dunn, 1989), a comprehensive learning styles instrument that includes perceptual constructs, and speaking and reading proficiency.

Bailey, Onwuegbuzie, and Daley (2000) state with regard to learning styles in general, rather than perceptual learning styles specifically

> There appears to be a gap in recent research between logical analyses of the importance of learning styles for foreign language learning and statistical confirmation of learning style preference as a direct measure of foreign language achievement. (p. 128)[7]

In the same study, and in an effort to address the deficit, Bailey et al. (2000) examined the role of a range of learning styles, including perceptual learning styles, in predicting foreign language achievement. The instrument of choice for these authors was the PEPS (Productivity Environmental Preference Survey), referred to above and a progenitor of the PLSPQ, which measures learning style in four major areas: preferences for environmental stimuli

(sound, light, etc); emotional stimuli; sociological stimuli; and physical stimuli (the category into which the auditory, visual, and kinesthetic modes of perception fall). Reliability for the subscales used in the study could not be assessed because scoring was done by the owners of the instrument (Bailey et al., 2000)—an extraordinary limitation to place on an instrument by any author or publisher. Foreign language achievement was measured using standardized course averages to accommodate for differences in teacher characteristics. Multiple regression (All Possible Subsets: APS) and correlation analysis were conducted with the learning styles constructs functioning as the independent variable and the achievement scores as the dependent variable. Of the perceptual learning styles constructs represented in the instrument, only kinesthetic preference correlated significantly with the dependent measure (achievement scores) accounting for approximately 4% of the variance. The auditory and visual constructs correlated very weakly with the dependent measure. In the APS multiple regression model, kinesthetic preference featured along with the design, responsibility, and mobility constructs. The overall model accounted for 15.1% of the variance.

In a less well-known study, Thomas, Cox, and Kojima (2000) obtained similar results. Achievement measures were TOEIC (Test of English for International Communication) scores and class grades, and the learning styles instrument was the PLSPQ. No correlation between styles and TOEIC scores was found, and only the kinesthetic scale (similar to the hands-on scale in the SAS) showed some correlation with course grades in practical, skills-based courses (the achievement measure). The kinesthetic (hands-on) construct is difficult to locate as a perceptual construct because much of the behavior associated with such a construct involves integrated perception, although tactile perception might be more dominant in the construct. Nonetheless, the two most salient modalities of perception, auditory and visual, do not feature in the results.

These results are modest to poor overall, and most importantly, provide no evidence for the predictive power of preferences for the most obvious perceptual constructs, namely, the visual and auditory modalities. It is also notable that these are the constructs which have proved the most problematic from a psychometric point of view (Table 1). Bailey et al. (2000), in accounting for the weak performance of learning styles constructs within their study, commented on instrumentation issues in their conclusion. While made in the context of learning styles in general, these comments are just as applicable to the subset of perceptual learning styles. The two most pertinent are a call for more situation-specific instruments, and an endorse-

ment of Skehan's (1991) recommendation that it may be necessary to move beyond questionnaires—a recommendation that has still not seen adoption; see more detailed discussion below. Suffice to state at this point that issues of instrumentation necessarily precede issues of demonstrating predictive power empirically, and that instrumentation issues may very well go to the heart of the disjunction between the intuitive appeal of perceptual learning styles constructs and the lack of empirical support for their usefulness. The track record of perceptual learning styles instrumentation is not good, and I argue that editorial oversight in enforcing the counsel of the APA, counsel specifically offered by Wilkinson and the APA Task Force on Statistical Inference (1999) and relating to the entry of defective measures into the literature, is critical.

Some Recommendations Related to Editorial Oversight on Entry of Instrumentation Into the Literature

A broader, noteworthy implication of the above analysis is that weak instruments have gained traction and persistence within applied linguistics contra the above-mentioned warnings of Wilkinson and the APA Task Force on Statistical Inference (1999). One lesson which should be learned from this state of affairs is that editorial oversight should be particularly austere with respect to new instruments entering the literature—particularly when these instruments are recommended to practitioners for classroom diagnostics. It is considerably harder to call attention to the limitations of an instrument after entry into the literature than it is to simply prevent entry right from the beginning. Once researchers are using an instrument, other researchers tend to assume that the instrument must be credible due to a mainstream effect. This is human behavior.

More specifically, and ideally,[8] I argue that new instruments should enter the literature via a dedicated psychometric study. By "dedicated," it is meant that the entire study focuses in a comprehensive way on the psychometric properties of scores produced by the instrument for an intended population, and the psychometric issue is not something that is simply auxiliary to other research goals.[9] In such a study, the standard quotation of Cronbach's alphas (an index that has come to be seen erroneously as a panacea for assessing reliability within much of applied linguistics) is not sufficient for reasons that will be outlined below. Insofar as demonstrating unidimensionality of scores is one important component of such a comprehensive study, CFA (with further discussion below) is advocated as a powerful tool in satisfying this requirement. It is important that my assertion of the negative case, that

alpha is not a panacea, not be read as implying a corresponding positive case that CFA is a panacea. CFA is a subset of structural equation modeling (SEM) and refers to the measurement model as opposed to the structural model. In other words, CFA is concerned with the manner in which observable measures reduce to purported underlying constructs rather than with how these constructs relate to each other. If a dedicated measurement paper is not within the remit of a particular applied linguistics journal, then the editor, assisted by the editorial board, should ensure that such a paper has been published or at least gone to press in a journal receptive to such studies before allowing the instrument to enter the literature for any other purpose via their own journal. Omission in this respect may involve the journal in being the conduit for an inadequate instrument gaining traction in the research community, and importantly contra the recommendations of Wilkinson and the APA Task Force on Statistical Inference (1999).

It was stated above that it is not sufficient to simply report Cronbach's alphas in the kind of dedicated psychometric study suggested. It needs to be emphasized that Cronbach's alpha is a minimal index and it is far from the state of the art. In fact, recent argumentation has drawn attention to its limitations and called for its replacement as a routine index of reliability (Bentler, 2009; Green & Yang, 2009a, 2009b; Revelle & Zinbarg, 2009; Sijtsma, 2009a, 2009b). Cronbach's alpha has historically been seen as useful in assessing reliability of scores in ongoing uses of an instrument for research purposes after a dedicated and comprehensive psychometric study for the instrument at launch time. This routine engagement with the reliability of scores is advisable because prior generation of reliable scores in one population does not guarantee scores with similar reliability in another population using the same instrument, and alpha has been seen as an easy and convenient reliability index to compute in dealing with this (though its continued suitability is a matter of objection as the citations above indicate). Wilkinson and the APA Task Force on Statistical Inference (1999) have emphasized that reliability is a property of scores and not instruments, and for this reason indexes of reliability should be reported for any new set of scores even when the purpose of a particular study is not psychometric.

The first limitation with respect to alpha has been pointed out by Cortina (1993), a study cited by Sijtsma (2009a), who drew on a Monte Carlo study[10] conducted by Green, Lissitz, and Mulaik (1977)[11] which was also a study cited by Sijtsma. This limitation is that Cronbach's alpha is partly a function of the number of items in a scale. In illustrating this point, Cortina compared the meaning of standardized alpha = .80 for two hypothetical scales of 3 and 10 items, respectively. Given this alpha level, the inter-item correlation for

the 10-item scale is .28 and for the 3-item scale .57. As Cortina pointed out, these inter-item correlations are strikingly different. This means that scales with a larger numbers of items are predisposed towards higher alpha levels (assuming inter-item correlations are constant) and that alpha levels should always be interpreted critically. More specifically, this has two implications. First, frequently quoted cut-off criteria for alpha, the most common one being Nunnally and Bernstein's (1994) criterion of .70, need to be considered critically when being applied, assuming the advisability of using alpha which is in dispute. Such criteria are useful rules-of-thumb but are not infallible. Second, the index should be understood and interpreted rather than simply reported. It is a diagnostic with limitations and not a panacea for detecting reliability.

Another limitation of alpha is that it is not a good measure of unidimensionality[12] which is a property of scores generated by a scale that needs to be empirically demonstrated if the scale is to be confidently interpreted. Alpha is more precisely a measure of reliability, which is not an equivalent property to unidimensionality. Cortina (1993) stated the following with regard to the psychological literature at the time, and there is a strong case that this pertains to much of applied linguistics literature today:

> The problem is that, just as the psychological literature reflects no clear understanding of the extent to which alpha is affected by the number of items, so does it reflect no clear understanding of the extent to which alpha is affected by dimensionality. (p. 101)

The intuitive assumption upon which many researchers are currently operating within aspects of applied linguistics is that if alpha is high, then this is evidence of unidimensionality for scores generated by the scale in question. This is not the case, and Cortina (in engaging with psychological literature) again drew on the Monte Carlo study of Green et al. (1977) to demonstrate this. Miller (1995) put the issue succinctly in stating that the proper use of alpha assumes unidimensionality rather than demonstrating it. It is entirely possible to arrive at a high alpha coefficient from a multidimensional scale. This is often readily apparent in the errant manner in which alpha is often reported with regard to instruments within applied linguistics literature—and here the review extends the claim to areas beyond learning styles. For example, Oxford (1996) reports very high alphas (above .90) for the Strategy Inventory for Language Learning (SILL) which comprises six subscales—and which has found significant use within contemporary ap-

plied linguistics research. In this case, high alphas are reported for an entire instrument that is not unidimensional and is, in fact, multidimensional by design.[13] And on a further note, the value for alpha in these quotations for entire multiscale instruments (involving many items) is often very high, and this can be expected given alpha's positive bias for number of items. Researchers would be better served by seeing alpha as assuming unidimensionality rather than demonstrating it; that is, if the index continues to be used in spite of the calls for its replacement.

With regard to this paper's advocacy of CFA as an important method for attending to the issue of unidimensionality of scores in comprehensive psychometric studies to attend the launch of new instruments, an influential article by Gerbing and Anderson (1988) made a case that CFA is the only method properly equipped to assess the unidimensionality of scales. I do not seek to endorse this strong version of the advocacy of CFA and its exclusiveness for testing unidimensionality, but do wish to draw the attention of the reader to its usefulness as one powerful instrument for this purpose. The goals of any researcher releasing new instrumentation into applied linguistics literature should, amongst other things, include demonstrating that scales making up the instrument generate scores that are indeed unidimensional. Subscales that produce multidimensional scores should be considered suspect in terms of validity because they cannot obviously, and necessarily, be interpreted in terms of the single label that semantically characterizes each scale.

Why CFA is powerful with respect to demonstrating unidimensionality is helpfully understood against the more frequently encountered method of EFA, which also engages with dimensionality, but which does not offer the same prospects for a direct test of a hypothesized unidimensional measurement model for an instrument. This applies whether the model in question is explicitly hypothesized by the author or implicitly hypothesized by virtue of the scoring regime offered along with the instrument. In EFA, as Thompson (2004) points out, simple structure is arrived at through a linear sequence of decisions which include:

1. Which matrix of association coefficients should be analyzed?
2. How many factors should be extracted?
3. Which method should be used to extract the factors?
4. How should the factors be rotated?
5. How should factor scores be computed if factor scores are of interest? (p. 27)

As Thompson and Daniel (1996, p. 204), in an earlier contribution, point out, an expected model either emerges out of this sequence or it does not, and rival models are not tested. Also, the decision of how many factors to extract, while not arbitrary, is attended by difficulties of determinacy. For example, the frequently-used, eigenvalue-greater-than-one rule (Guttman, 1954; Kaiser, 1961) can overestimate the number of factors (Zwick & Velicer, 1986) and inspection of a scree plot (Cattell, 1966) is perceptually subjective. In CFA, the researcher approaches the data set with an a priori model which is tested directly against the data. The unidimensional model is specified so that observables (measured items on the instrument itself) load only on the factor they are hypothesized to indicate and not on other factors and this is unlike EFA, wherein observables can indicate all factors. Adjudication of the fit of the a priori model is conventionally undertaken using a variety of indexes such as the standardized root mean square residual (SRMSR), the root mean square error of approximation (RMSEA), the comparative fit index (CFI) and the Tucker-Lewis index (TLI). There are two important points to note with regard to the use of these indexes. First, the fact that a variety is used provides for triangulation of the decision as to whether the model is satisfactory or not. Second, the cut-off criteria conventionally used to adjudicate model fit (see, for example, Hu & Bentler, 1999) are empirically derived to minimize both Type I and Type II error. This assists, in an evidential way, with the problem of determinacy in adjudicating model fit. Finally, and unlike EFA, the testing of an a priori model can be conducted in the context of testing rival models. Even if the a priori model fits satisfactorily, it is, therefore, possible to test whether other plausible models fit better. This kind of analytical leverage exceeds that available to EFA, and the explanation here is cursory (for further explanation and understanding see Byrne, 2001, 2005; Schmitt, 2011; Thompson, 2004).

Turning away from the statistical concerns referred to above, one of the issues that has emerged in perceptual learning styles research is foreign language instrumentation. A precedent for administration of perceptual learning styles instruments in a language foreign to the respondents was set by Reid (1987), who used an English-language version for NNSs. This precedent has been followed in a number of studies (Bowman, 1996; Peacock, 2001; Rossi-Le, 1995; Stebbins, 1995), but not all studies, leading up to the present. As recently as 2005, DeCapua and Wintergerst have defended their use of English-language instrumentation for NNSs in a line of research using the PLSPQ (or new version of it) that includes four studies (DeCapua & Wintergerst, 2005; Wintergerst & DeCapua, 2001; Wintergerst et al., 2002,

2003). DeCapua and Wintergerst (2005) cited Eliason (1995) who cited Melton (1990) and Inclan (1986) in making an unassertive case for foreign language instrumentation. Eliason reported that Melton's study found no significant difference in scores (using ANOVA and Tukey's Multiple Comparisons of Means) for individuals when the instrument was administered in both English and Chinese. The reasoning here for the case of equivalence of measurement was weak. The psychometric case for different language versions of an instrument measuring equivalently has to be made on the specific merits of the scores generated by each language version of the instrument, in the population for which it was translated, using EFA, CFA,[14] and other methods related to psychometrics as the methodology, and not through the use of standard inferential statistics involving samples of convenience to see if one arrives at a nonsignificant result. This approach taken by Melton was unorthodox,[15] was not the main part of his study, and was no foundation for a precedent.

The citation of these two particular studies (Inclan, 1986; Melton, 1990) by Eliason (1995) and DeCapua and Wintergerst's subsequent (2005) citation of Eliason indicate a view on the part of these authors that some kind of empirical case can be made for the use of language versions of instruments that are foreign to the respondent. However, there is a far stronger case that instruments in language versions foreign to the respondent become progressively more untenable as the language competence of the respondent decreases. An English-language version of the PLSPQ, or any other instrument for that matter, might seem to function with students above a certain threshold (which respondents in DeCapua and Wintergerst's study might have been), but any instructor with the experience of facing classes of beginners will be unconvinced by foreign language instrumentation. An empirical case hardly needs to be made for this. If researchers want instrumentation to accommodate all levels of students, they should translate instruments into the native language of the intended respondents. This is a criterion which needs to be explicitly adopted in any editorial review processes within applied linguistics journals.

It is also notable that this trend toward using foreign language instrumentation within aspects of applied linguistics, which I would argue is significant rather than pervasive, is contrary to the Test Adaptation Guidelines of the International Test Commission (ITC, 2001). The Commission is extensively engaged with establishing statistical and methodological guidelines for adapting tests across languages and cultures, and the precedent for foreign language instrumentation set in the perceptual learning styles research

within applied linguistics directly contradicts this. Any exercise of editorial prerogative to exclude foreign language instrumentation from applied linguistics research and journals would be consistent with the guidelines of the ITC.[16]

A Revised PLSPQ and the Future of Instrumentation in Perceptual Learning Styles

As part of this critique of perceptual learning styles research, I now turn to the Learning Styles Inventory (LSI), which is important because it is a recent incarnation of the PLSPQ which may have an impact on the future of learning styles research. The instrument represents an effort to revise the PLSPQ in view of the evident problems, but I argue that the approach, nonetheless, inherits these problems and that the new instrument is premised on a misunderstanding of the concept of reliability via its use of repetitive items.

The LSI is essentially a residual-item PLSPQ under a new hypothesized structure after a diagnostic EFA published by Wintergerst et al. (2001). It has appeared in four other studies (DeCapua & Wintergerst, 2005; Wintergerst & DeCapua, 2001; Wintergerst et al., 2002, 2003). While it represents the latest generation of instrumentation in perceptual learning styles research, the LSI is conspicuous for having abandoned the most salient perceptual constructs (visual and auditory perception) in this process of revision. The hypothesized three-construct model includes three explicit scales: Group Activity Orientation (GAO), Individual Activity Orientation (IAO), and Project Orientation (PO). These scales incorporate most, but not all, of the original items from the PLSPQ.

I question the appropriateness of retaining the items making up the PLSPQ under a different hypothesized structure. The theoretical rationale for the new hypothesized structure for the PLSPQ, in the form of the constructs in the LSI cited above, came after the fact of the EFA rather than before it. A better and more scientifically credible procedure, having established that the PLSPQ generates psychometrically weak scores, would be to proceed with an entirely new instrument based on a priori theoretical reasoning. The hypothesized constructs emerging from such a theoretical rationale should employ, and benefit from, a large and diverse exploratory set of items targeted at the hypothesized constructs rather than the limited set of items in the PLSPQ. The items in the PLSPQ were designed with a different theoretical rationale and hypothesized structure in mind and represent a

post-reduction set of items (reduced by Reid in the original development of the PLSPQ). Furthermore, they were simplified, often to the point of repetitiveness and paraphrasing one another (an issue covered in more detail below), to accommodate the nonnative speaker (Reid, 1990)—a process which belies the view that foreign language instrumentation is not really problematic.

I argue that the line of research being conducted by DeCapua and Wintergerst (2005) with the LSI inherits and perpetuates the problems of the past. The PLSPQ is a very problematic instrument by the author's own admission (Reid, 1990), and it is clearly time to abandon this artifact of applied linguistics' psychometric legacy.

Turning to the issue of repetitive questions in the LSI (inherited from the PLSPQ), DeCapua and Wintergerst (2005) state the following with respect to their use of the instrument in one study:

> Both in the class discussions and in the interviews, informants pointed out repeatedly the repetition among the statements. The students were very much aware that the same questions were asked in different ways on the LSI and questioned why. Indeed, in the class discussion several students mentioned that they had thought this was some sort of mistake. Even though the instructor pointed out that stating the same thing in more than one way is a way of ascertaining whether there is consistency in responses, the students still felt that this was a weakness of the LSI. (p. 9)

The informants rather than the instructor have the stronger case here. The purpose of having more than one item measuring a construct is not to ascertain consistency in responses if that means repeating the same question to some degree or another. The purpose is rather to allow for a more diverse and exhaustive operational expression of the underlying construct. If a scale comprises many diverse items, the idiosyncrasies of each specific item in measuring the respective construct will be averaged out. Statements should not be repeated or paraphrased to see if there is consistency in response, because the consistency or stability of the item is with respect to the measurement of the underlying construct and not the item itself. Repeating the item simply repeats the idiosyncrasy of that specific item. Statements that are virtual paraphrases of each other often inter-correlate highly and produce higher alphas but the researcher should hardly be surprised by this. If you ask someone the same question twice, you should not be surprised if you

get the same answer. This use of paraphrases produces measured constructs that have been referred to by Kline (1994) as "bloated specifics." In addition, the cost of this artificially high internal consistency is that construct bandwidth is sacrificed because there is really only one operational expression of the construct repeated many times. The goal should be to reach optimum levels of both internal consistency and construct bandwidth, and this cannot be done with the items in the PLSPQ, which are flawed as a result of their repetitiveness.

The Future of Perceptual Learning Styles

Perceptual learning styles represent a special case for measurement due to their resistance to operationalization—especially through self-report. Self-report assumes metacognitive awareness of what is being reported and this assumption is not always satisfied. Most of what we perceive as humans is an integrated perceptual experience, and it may be quite difficult for respondents to distil this integrated experience into pure perceptual modalities and make an authentic judgment on which one they favor.

The view that perceptual preference is resistant to operationalization is supported by early research within education (Deverensky, 1978; Dunn, 1990; Kampwirth & Bates, 1980; Kavale & Forness, 1987, 1990; Tarver & Dawson, 1978) and a stream of research experience within applied linguistics that is very much a facsimile of this earlier educational research. There is also an interesting reprise of this issue, again from outside the field of applied linguistics, where operational issues and the usefulness of the constructs have been empirically challenged (Kratzig & Arbuthnott, 2006). Interestingly, this resistance to operationalization is a critical feature of the LSI's emergence out of the PLSPQ that is not explicitly referred to by DeCapua and Wintergerst (2005). The PLSPQ was originally a perceptual learning styles instrument (Visual, Auditory, Kinesthetic, and Tactile) including two nonperceptual constructs (Group and Individual). The LSI, as a revised version of the original instrument, is a nonperceptual learning styles instrument including, essentially, the original Group and Individual constructs (now the GAO and IAO) and the new Project Orientation (PO) construct, which is an amalgam of the Tactile and Kinesthetic items. This means that in revising the PLSPQ, the notion of perceptual learning styles in the instrument has effectively been abandoned. This fact is very revealing and is not surprising because research into the PLSPQ has consistently shown that the perceptual constructs are weak and the nonperceptual constructs strong (Isemonger & Sheppard, 2007; Itzen, 1995; Wintergerst et al.,

2001). It is clear that perceptual constructs are not viably measured through self-report.

Bailey et al. (2000) endorsed Skehan's (1991) call to move beyond questionnaires with respect to learning styles. With regard to the preference for perceptual learning styles, specifically, this is particularly the case. None of the instruments so far have succeeded in producing valid scores. This recommendation, however, has different implications for practitioners and researchers. For the researcher, there may well be methods within a laboratory setting to reliably determine preferences for perceptual modality. However, for the practitioner, the self-report method is pervasive because it is convenient, efficient, and unobtrusive. If preference for perceptual learning styles cannot reliably be determined using this method, then there may well be no way to effectively measure such styles in the classroom. Given that the literature has yet to credibly demonstrate the predictive power of perceptual learning styles constructs in achievement terms, I argue that practitioners need not lose any sleep over this. There are other instruments in other areas of individual differences that present far better prospects for pedagogical intervention.

Conclusion

It is important to reiterate that editorial oversight in the area of psychometrics is critical to avoiding the problems that have attended the perceptual learning styles research trajectory within applied linguistics. Admitting questionable instruments into the literature is something that must be vigilantly guarded against. I do recognize the difficulties editors face in having to work with the submissions they have rather than the submissions they would desire, and perhaps this also raises the importance of appropriate training in postgraduate applied linguistics programs. What is clearly arguable is that the burden of improving oversight with respect to these issues falls first to leading journals which set the precedent for the field and enjoy greater freedom in their editorial decisions for having more and better submissions.

The conclusions from this critical review are also significant for teachers and practitioners, who have limited time for diagnostics related to individual differences in learning outcome—they do after all have to get down to the core task which is to teach language. There are instruments in other areas of individual differences which have less questionable psychometrics and a far better track record in predicting learning outcome (e.g., anxiety and motiva-

tion), though I am not at all claiming that deficiencies I have pointed out are exclusive to the perceptual learning styles domain in applied linguistics. I would not have cited the case of perceptual learning styles as instructive if the deficiencies were unique. For teachers and practitioners who remain concerned about perceptual preferences, it is arguable that paying attention to offering a good multimodality class covers the proclivities of all groups of students minus the fuss of having to diagnose what preference each student has. Furthermore, it is arguable that all groups are better served by an integrated, multimodality learning experience than a singular modality experience which caters to the preference of a particular group—with all the attendant problems of coping with the groups that might be marginalized by such disadvantageous matching. Assuming that preferences for perceptual modality do exist, even if difficult to detect, there is no reason at all to go on to assume that a visually oriented student will perform better in a visually delivered class than in a multimodality class. In fact, it could be predicted that all perceptual orientations will perform better in an integrated multimodality class, and this prediction might be tested in future research if such orientations or preferences can be reliably detected.

Finally, practitioners should not assume that psychometric instruments (measuring any kinds of constructs) which come out of published literature produce valid scores simply because such instruments have been published and have been used by others. Practitioners should check for good psychometric studies that thoroughly question an instrument, and should not simply be satisfied with uncritical quotations of the value for Cronbach's alpha. High values for alpha quoted for an entire instrument that is multidimensional by design are not very helpful because interpretation usually occurs at the level of each dimension or scale, and anyway alpha does not effectively demonstrate unidimensionality, which is a critical property for score interpretation. Furthermore, high values for alpha on scales which contain highly paraphrased items should be treated with considerable skepticism. In terms of all of the above, the line of instrumentation within applied linguistics measuring perceptual learning styles is questionable and practitioners should avoid using these diagnostics in class until the instruments are shown to be capable of generating valid scores, and until there is serious evidence for the predictive power of the constructs they attempt to measure in terms of learning outcome. There is an opportunity cost in running diagnostics in language class, and this cost is lost time for the core business of language teaching. This opportunity cost should only be borne if the benefits from the diagnostics have been empirically demonstrated to be significant.

Notes

1. I also neglected this work as a new researcher entering this area (Isemonger & Sheppard, 2003). In the same study in 2003, I also failed to observe some of the recommendations made in this paper. I have also pointed out these failures once before (5 years ago) in Isemonger and Sheppard (2007). In addition, some of the claims about perceptual learning styles in the 2003 paper illustrate how much I have changed my mind since that time.
2. I have recounted this history of learning styles controversy which occurred outside of applied linguistics once before in an article in the *Journal of Psychoeducational Assessment* (Isemonger, 2008).
3. In a meta-study drawing on a sample of 696 tests appearing in the APA-published Directory of Unpublished Experimental Mental Measures (Hogan, Benjamin, & Brezinski, 2000), 85% of reported alpha coefficients met this criterion of .70. However, as will be argued later in this paper, Cronbach's alpha is an index that should be interpreted. The number of items on a scale "biases" (positively) the value derived for the index as true score variance builds up faster than error variance as items are added. In the case of the results presented for the Auditory and Visual Scales on the PLSPQ, there is little need for interpretation of alpha with respect to this bias since the value for alpha is so low.
4. CFA is a more sophisticated and appropriate verificational tool than EFA and also allows for the direct testing of the superiority of rival structures for an instrument. There is more discussion on this later in the paper.
5. Again, and this will be explained in more detail further on in the paper, Cronbach's alpha is biased (positively) by the number of items in a scale: 10 to 12 items would be considered a fairly high number of items for a scale leading to an expectation of higher alphas.
6. CFA is a method that has been adopted more recently than EFA due to the specialized software required for it, and retrospective criticism with regard to the absence of its use for emerging instrumentation should be understood in this context.
7. Bailey, Onwuegbuzie, and Daley do cite a study conducted by Oxford et al. (1993) which dealt with perceptual learning styles using the LCPC and which claimed that the Visual construct was "more predictive of Japanese language achievement" (p. 367) than the Auditory and Kinesthetic constructs. However, no achievement data is presented in the Oxford et al. study, nor were the statistical methods for arriving at this conclusion reported. No models were presented indicating the respective amounts

of variance accounted for by these constructs vis-à-vis other constructs examined in the same study including motivational and learning-strategy constructs. Furthermore, alphas for the LCPC subscales were poor given that all fell well below .70 on 12-item scales. Bailey, Onwuegbuzie, and Daley also review some other studies that consider the predictive power of learning styles, but the studies reviewed examine nonperceptual learning styles constructs.

8. I say "ideally" here because journal editors deal with the reality that the standard of current practices for the submissions they receive (which is of course highly contingent on the stature of the journal) may not necessarily be conducive to enforcing the best practice. Ultimately, the burden of implementation should fall first on the leading journals in the expectation that methodological practices in such journals set the precedent and cascade down to journals with harder choices to make at press time.
9. If the instrument produces valid scores in the first dedicated study, then this should also be followed up with further similar studies in all populations where it is intended for use. Evidence in this regard is a cumulative process.
10. A Monte Carlo study is a study in which certain known properties are built into a data set.
11. I realize that some of these citations may seem dated, but the uncritical manner in which alpha is used as an index of reliability within applied linguistics indicates that their central point has yet to be taken on board. More recent citations of these older studies by Sijtsma indicate that this is also the case in other research fields, and that these older citations are still pertinent.
12. In simple terms, unidimensionality refers to the property that the scale measures quite *exclusively* what the author claims it measures, and does not measure other (possibly unknown) dimensions of behavior that may complicate the interpretation of the scale.
13. The rationale for this reporting of a composite alpha for the entire SILL is not clear. If a second order CFA model hypothesized a superordinate construct which subsumed the six subscales and this model was confirmed and worked as well, then there would be a rationale for quoting alpha for the entire instrument, because interpretation of a composite score for the six subscales would make sense. This might occur with constructs such as anxiety where different facets of anxiety constitute subcomponents of a more generalized anxiety disposition. I do not see

how it occurs in the case of the SILL, and certainly no case has been made for a superordinate and interpretable construct for the overall SILL. Alpha is only meaningful at the level at which interpretation takes place—and unidimensionality should have been previously demonstrated at this level for scores derived from the population for which the scale is being used.

14. One of the powerful ways to establish equivalency of measurement across translated versions of an instrument (administered in their respective populations) is to use the group analysis capabilities of CFA in what is often referred to as a measurement invariance study. Byrne (2001) offers helpful chapters on this analytical procedure.
15. A new language version of an instrument is essentially a new instrument for a new population. Its equivalence with the original needs to be empirically demonstrated and CFA is an appropriate method for this.
16. Quite obviously, this would not include instrumentation where the foreign language being learned is the subject of the test.

Ian Isemonger is an Associate Professor at Kumamoto University where he teaches in the Department of Communication and Information Studies and in the Graduate School's TESOL master's and doctoral programs.

References

Bailey, P., Onwuegbuzie, A. J., & Daley, C. E. (2000). Using learning style to predict foreign language achievement at the college level. *System, 28*, 115-133.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*, 137-143.

Bowman, A. (1996). ESL students learning style preferences in a multimedia language laboratory: Do students do what they say they do? *University of Hawaii Working Papers in ESL, 15*, 1-31.

Byrne, B. M. (2001). *Structural equation modeling with AMOS.* London: Lawrence Erlbaum Associates.

Byrne, B. M. (2005). Factor analytic models: Viewing the structure of an assessment instrument from three perspectives. *Journal of Personality Assessment, 85*, 17-32.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104.

Deaton, W. L. (1992). Buros review of LCPC. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

DeCapua, A., & Wintergerst, A. C. (2005). Assessing and validating a learning styles instrument. *System, 33*, 1-16.

Deverensky, J. L. (1978). Modal preferences and strengths: Implications for reading research. *Journal of Reading Behaviour, 10*, 7-23.

Dunn, R. (1983). Learning style and its relation to exceptionality at both ends of the spectrum. *Exceptional Children, 49*, 496-506.

Dunn, R. (1984). Learning style: State of the scene. *Theory into Practice, 23*, 10-19.

Dunn, R. (1990). Bias over substance: A critical analysis of Kavale and Forness' report on modality-based instruction. *Exceptional Children, 56*, 352-356.

Dunn, R., & Dunn, K. (1972). *Practical approaches to individualizing instruction*. Englewood Cliffs, NJ: Prentice Hall.

Dunn, R., & Dunn, K. (1979). Learning styles/teaching styles: Should they ... can they ... be matched? *Educational Leadership, 36*, 238-244.

Dunn, R., Dunn, K., & Price, G. E. (1975). *The Learning Style Inventory*. Lawrence, KS: Price Systems.

Dunn, R., Dunn, K., & Price, G. E. (1978). *Teaching students through their individual learning styles*. Reston, VA: Reston Publishing.

Dunn, R., Dunn, K., & Price, G. E. (1979). *The Productivity Environment Preference Survey*. Reston, VA: Reston Publishing.

Ehara, Y. (1998). Differences in learning styles between Japanese teachers of Japanese and Mexican learners of Japanese. *Journal of the Society for Teaching Japanese as a Foreign Language, 96*, 13-24.

Ehrman, M., & Oxford, R. L. (1995). Cognition plus: Correlates of language learning success. *Modern Language Journal, 79*, 67-89.

Eliason, P. A. (1995). Difficulties with cross-cultural, learning-styles assessment. In J. M. Reid (Ed.), *Learning styles in the ESL/EFL classroom* (pp. 19-33). Boston, MA: Heinle & Heinle.

Frank, M., & Hughes, M. (2002). Examining the learning styles of Japanese students at Keiwa College. *Keiwa Gakuen University Research Bulletin, 11*, 73-86.

Gerbing, D. W., & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research, 15*, 186-192.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*, 827-838.

Green, S. B., & Yang, Y. (2009a). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74*, 121-135.

Green, S. B., & Yang, Y. (2009b). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*, 155-167.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika, 19*, 149-161.

Henry-Vega, G. (2004). *Exploratory study on the processing styles and the processing strategies of 2 second language graduate students when reading texts for academic purposes.* University of Cincinnati, Cincinnati, OH.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60*, 523-531.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Hughes, M. (2001). Examining the learning styles and learning strategies used by Japanese learners of English. *Keiwa Gakuen University Research Bulletin, 10*, 59-87.

Hyland, K. (1993). Culture and learning: A study of the learning styles of Japanese students. *RELC Journal, 24*, 69-91.

Inclan, A. F. (1986). The development of the Spanish version of the Myers Briggs type indicator, form G. *Journal of Psychological Type, 11*, 35-46.

International Test Commission. (2001). *International Test Commission guidelines for test adaptation*. London: Author.

Isemonger, I. M. (2008). Scores on a Japanese-language version of the Learning Channel Preference Checklist: A questionable instrument within a questionable line of instrumentation. *Journal of Psychoeducational Assessment*, *26*, 148-155.

Isemonger, I. M., & Sheppard, C. (2003). Learning styles. *RELC Journal, 34*, 195-222.

Isemonger, I. M., & Sheppard, C. (2007). A construct-related validity study on a Korean version of the Perceptual Learning Styles Preference questionnaire. *Educational and Psychological Measurement, 67*, 357-368.

Isemonger, I. M., & Watanabe, K. (2007). The construct validity of scores on a Japanese version of the perceptual component of the Style Analysis Survey (SAS). *System, 35*, 134-147.

Itzen, R. (1995). *The dimensionality of learning structures in the Reid Perceptual Learning Style Preference Questionnaire* (Unpublished doctoral dissertation). Graduate College of the University of Illinois at Chicago.

Kaiser, H. (1961). A note on Guttman's lower bound for the number of common factors. *Multivariate Behavioral Research, 1*, 249-276.

Kampwirth, T. J., & Bates, M. (1980). Modality preference and teaching method: A review of the research. *Academic Therapy, 15*, 597-605.

Kavale, K. A., & Forness, S. R. (1987). Substance over style: Assessing the efficacy of modality testing and teaching. *Exceptional Children, 54*, 228-239.

Kavale, K. A., & Forness, S. R. (1990). Substance over style: A rejoinder to Dunn's animadversions. *Exceptional Children, 56*, 357-361.

Keefe, J. W., Monk, J. S., Letteri, C. A., Languis, M., & Dunn, R. (1989). *Learning style profile*. Reston, VA: National Association of High School Principals.

Kelly, C. A. (1998). The learning style preferences of Japanese EFL students. *Gakuen, 697*, 30-36.

Kim, H. J. (2001). Language learning strategies, learning styles and beliefs about language learning of Korean university students. *Journal of the Pan-Pacific Association of Applied Linguistics, 5*, 31-46.

Kinsella, K. (1995a). Perceptual Learning Preferences Survey. In J. M. Reid (Ed.), *Learning styles in the ESL/EFL classroom* (pp. 221-238). Boston, MA: Heinle & Heinle.

Kinsella, K. (1995b). Understanding and empowering diverse learners in the ESL classroom. In J. M. Reid (Ed.), *Learning styles in the ESL/EFL classroom* (pp. 170-193). Boston, MA: Heinle & Heinle.

Kline, P. (1994). *An easy guide to factor analysis*. London: Routledge.

Kratzig, G. P., & Arbuthnott, K. D. (2006). Perceptual learning styles and learning proficiency: A test of the hypothesis. *Journal of Educational Psychology*, *98*, 238-246.

Melton, C. D. (1990). Bridging the cultural gap: A study of Chinese students' learning style preferences. *RELC Journal, 21*, 29-47.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling*, *2*, 255-273.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

O'Brien, L. (1990). *Learning Channel Preference Checklist (LCPC)*. Rockville, MD: Specific Diagnostic Services.

O'Brien, L. (2002). *The Learning Channel Preference Checklist (LCPC)*. Kensington, MD: Specific Diagnostics Services.

O'Donoghue, R. K., Oyabu, T., & Akiyoshi, R. (2001). An exploratory survey of Japanese EFL students' preferred learning styles. *Bulletin of Fukuoka Dental College, 28*, 9-17.

Oxford, R. L. (1993a). *Style Analysis Survey (SAS)*. Tuscaloosa, AL: University of Alabama.

Oxford, R. L. (1993b). *The Style Analysis Survey (SAS) on CCET, University of Alabama*. Retrieved February 2010 from: http://www.as.ua.edu/nihongo/sas/survey.html

Oxford, R. L. (1995). Gender differences in language learning styles: What do they mean? In J. M. Reid (Ed.), *Learning styles in the ESL/EFL classroom* (pp. 34-46). Boston, MA: Heinle & Heinle.

Oxford, R. L. (1996). Employing a questionnaire to assess the use of language learning strategies. *Applied Language Learning, 7*, 25-45.

Oxford, R. L., & Anderson, N. J. (1995). A cross cultural view of learning styles. *Language Teaching, 28*, 201-215.

Oxford, R. L., Young, P. O., Ito, S., & Sumrall, M. (1993). Japanese by satellite: Effects of motivation, language learning styles and strategies, gender, course level, and previous language learning experience on Japanese language achievement. *Foreign Language Annals, 26*, 359-371.

Peacock, M. (2001). Match or mismatch? Learning styles and teaching styles in EFL. *International Journal of Applied Linguistics, 11*, 1-20.

Price, G. E., & Dunn, R. (1997). *The learning style inventory: LSI manual.* Lawrence, KS: Price Systems.

Price, G. E., Dunn, R., & Dunn, K. (1996). *Productivity environmental preference survey: PEPS manual*. Lawrence, KS: Price Systems.

Price, G. E., Dunn, R., & Sanders, W. (1980). Reading achievement and learning style characteristics. *The Clearing House, 5*, 223-226.

Psaltou-Joycey, A., & Kantaridou, Z. (2011). Major, minor, and negative learning style preferences of university students. *System*, *39*, 103-112.

Reid, J. M. (1987). The learning style preferences of ESL students. *TESOL Quarterly, 21*, 87-109.

Reid, J. M. (1990). The dirty laundry of ESL survey research. *TESOL Quarterly, 24*, 323-338.

Reid, J. M. (1995). *Learning styles in the ESL/EFL classroom*. Boston, MA: Heinle & Heinle.

Reid, J. M. (1998a). Teachers as perceptual learning styles researchers. In J. M. Reid (Ed.), *Understanding learning styles in the second language classroom* (pp. 15-26). Englewood Cliffs, NJ: Prentice Hall Regents.

Reid, J. M. (1998b). *Understanding learning styles in the second language classroom*. Englewood Cliffs, NJ: Prentice Hall Regents.

Reid, J. M. (2000). The perceptual learning style questionnaire. In J. M. Reid (Ed.), *The process of composition* (pp. 328-330). New York: Prentice Hall Regents.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the GLB: Comments on Sijtsma. *Psychometrika, 74*, 145-154.

Rossi-Le, L. (1995). Learning styles and strategies in adult immigrant ESL students. In J. M. Reid (Ed.), *Learning styles in the ESL/EFL classroom* (pp. 118-125). Boston, MA: Heinle & Heinle.

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessement, 29*, 304-321.

Shen, M. Y. (2010). Effects of perceptual learning style preferences on L2 lexical inferencing. *System*, *38*, 539-547.

Siew Luan, T.-K., & Ngoh, J. (2006). Innovative modes of continual assessment: Perspectives of undergraduate students. *Reflections on English Language Teaching, 5*, 47-64.

Sijtsma, K. (2009a). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120.

Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika, 74*, 169-173.

Skehan, P. (1991). Individual differences in second language learning. *Studies in Second Language Acquisition, 13*, 275-298.

Stebbins, C. (1995). Culture-specific perceptual-learning-style preferences of post secondary students of English as a second language. In J. M. Reid (Ed.), *Learning styles in the ESL/EFL classroom* (pp. 108-117). Boston, MA: Heinle & Heinle.

Tarver, S. G., & Dawson, M. M. (1978). Modality preference and the teaching of reading: A review. *Journal of Learning Disabilities, 11*, 17-29.

Thomas, H., Cox, R., & Kojima, T. (2000, March). *Relating preferred learning style to student achievement.* Paper presented at the Annual Meeting of the Teachers of English to Speakers of Other Languages, Vancouver, BC, Canada.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis.* Washington, DC: APA.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*, 197-208.

Wilkinson, L., & American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

Wintergerst, A. C., & DeCapua, A. (2001). Exploring the learning styles of Russian-speaking ESL students. *The CATESOL Journal, 13*, 23-46.

Wintergerst, A. C., DeCapua, A., & Itzen, R. C. (2001). The construct validity of one learning styles instrument. *System, 29*, 385-403.

Wintergerst, A. C., DeCapua, A., & Verna, M. A. (2002). An analysis of one learning styles instrument for language students. *TESL Canada Journal, 20*, 16-37.

Wintergerst, A. C., DeCapua, A., & Verna, M. A. (2003). Conceptualizing learning style modalities for ESL/EFL students. *System, 31*, 85-106.

Yamashita, S. (1995). The learning style preferences of Japanese returnee students. *ICU Language Research Bulletin, 10*, 59-75.

Yu-rong, H. (2007). A survey on the learning style preference of Tibetan EFL learners in China. *US-China Foreign Language, 5*, Serial No. 43.

Zwick, W. R., & Velicer, W. F. (1986). Factors influencing five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432-442.

Comparing the Story Retelling Speaking Test With Other Speaking Tests

Rie Koizumi
Tokiwa University
Akiyo Hirai
University of Tsukuba

This study examines the validity of score-based interpretation of the Story Retelling Speaking Test (SRST) in comparison with the Versant (Pearson Education, 2008) and Standard Speaking Test (SST; ALC Press, 2010). In total, 64 participants took the three tests; their speaking functions, scores, and utterances were analyzed to probe the shared and varied aspects of the tests. The results showed that the SRST elicited more functions than the Versant but fewer than the SST, that it was moderately related to the latter two tests, and that it more successfully discriminated among a group of beginner and intermediate level learners. Additionally, the results suggested that (a) the tasks and speaking functions and (b) the aspects emphasized while rating may explain the differences in the test scores for the three tests. Based on the results, comparative advantages of each test were summarized, which may be useful for selecting appropriate speaking tests according to assessment purposes and situations.

本研究では、Story Retelling Speaking Test (SRST) を、Versant (Pearson Education, 2008) と Standard Speaking Test (SST; ALC Press, 2010) と比較することで 、SRSTの得点に基づく解釈の妥当性を吟味する。64名の受験者に3つのテストを受けてもらい、テストの共通点と相違点を調べるために、その言語機能と得点、発話を分析した。その結果、SRSTはVersantよりは多いがSSTよりは少ない言語機能を引き出すこと、SRSTは他の2つのテストと中程度の相

関を持ち、初級者・中級者グループで弁別力を発揮すること、(a)タスクと言語機能と(b)評価時に重きを置く要素の相違により、テスト得点の違いが説明されうることが分かった。結果に基づき、評価の目的と状況に沿って適切なスピーキングテストを選ぶ際に有益となる、各テストの相対的な利点を示した。

One difficulty related to speaking tests is ensuring that their administration and scoring are sufficiently practical. This seems especially true when tests are undertaken for formative and summative classroom assessment. While teachers can observe students' class performance in pair and group interactive activities as well as speech, discussion, and other presentation activities, speaking tests are needed to grasp students' achievement and proficiency in relation to speaking ability (Genesee & Upshur, 1996). A classroom speaking assessment can take a direct (or live) test format, such as one-on-one interviews with a teacher and interactions with a partner or group members; however, difficulties may arise as to securing interviewers and having time for such direct testing. When equipment for recording students' voices is available, employing a semi-direct (or tape-mediated) format becomes a viable alternative to direct speaking assessment in which "the stimulus is pre-recorded or text based, and the response by the candidate is recorded for distance rating" (Davies et al., 1999, p. 178). One example of semi-direct tests is the Telephone Standard Speaking Test (TSST), in which test-takers talk about their experiences, describe objects, and compare two objects through telephone (ALC Press, 2008). Another example is the speaking component of the TOEIC® (Test of English for International Communication) Speaking and Writing Test; test-takers read a text aloud into a computer microphone, describe pictures, answer questions, propose solutions, and express opinions (Educational Testing Service, 2011). These examples illustrate that semi-direct speaking tests adopt several tasks to elicit various types of performance from test-takers. However, a semi-direct task that has hitherto been underutilized is story retelling, in which test-takers retell a passage that they have just read or heard. This integrated speaking activity simulates natural speech in real-life situations.

A tape-mediated Story Retelling Speaking Test (SRST) was developed for Japanese learners of English as a practical resource for classroom use to assess speaking ability, especially the ability to produce extended spoken monologues (Hirai & Koizumi, 2009). In the test, students read a story silently, then retell the story, and express their opinions about it while looking only at keywords (see Figure 1 for the procedure and Appendix A for instructions and a story sample). The administration of the SRST with

the test instructions and one story takes about 8 minutes. Utterances are recorded and rated using an empirically derived, binary-choice, boundary-definition (EBB) rating scale to assess three criteria with five levels each: Communicative Efficiency (CE; including fluency, coherency, elaboration, adequacy of story-telling capability, and aptness of test-takers' opinion of the story), Grammar & Vocabulary (G&V), and Pronunciation (see Appendix B for the EBB rating scale). The descriptors of the scale were empirically derived on the basis of previous literature (e.g., Upshur & Turner, 1995), and included the salient aspects of students' actual speech delivery that separate higher and lower levels of the EBB scale (Hirai & Koizumi, 2011). The use of three rating criteria on the EBB scale is intended to increase the diagnostic value of the score report.

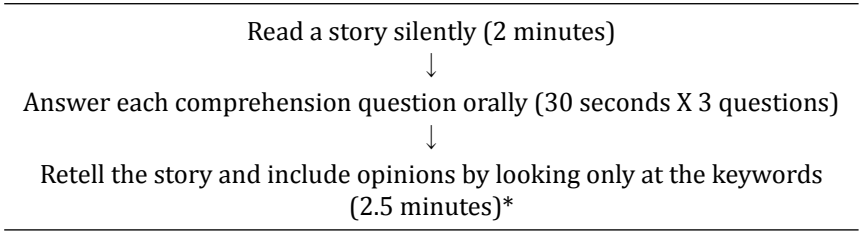


Figure 1. SRST Administration Procedure for One Story

*A beep sound is inserted after 2 minutes to inform test-takers of the time remaining (30 seconds) and when they should start expressing opinions.

The SRST is intended to have high practicality for test construction and administration: Teachers can convert lesson material into a test passage and the test procedure can be standardized with the recorded instructions. Although the test task is limited to story retelling and opinion statement, these skills are worth teaching and testing since L2 learners, especially Japanese learners of English, often lack skills in expressing their knowledge and opinions (National Institute for Educational Policy Research of Japan, 2007).

Previous studies (Hirai & Koizumi, 2009, 2011; Koizumi & Hirai, 2010) have examined test qualities of the SRST and shown evidence of its validity and usefulness. Hirai and Koizumi (2009) conducted a survey analysis and confirmed that test-takers generally felt that the test procedures and task difficulty were appropriate. In another study (Koizumi & Hirai, 2010), the effectiveness of the SRST components (e.g., keywords and opinions) was demonstrated by scrutinizing examinees' performances. For example, the

effect of text length on volume produced was found to be inconsistent and small, which suggests that memory has only a slight impact on SRST performance. Hirai and Koizumi (2011) compared two empirically developed rating scales (i.e., EBB vs. multiple trait) and demonstrated that the EBB scale has the more desirable characteristics of requiring fewer stories to maintain sufficient reliability (.70 or above) and of showing stronger discrimination. However, concerns regarding the validity of interpretation and use of the SRST scores remain. For example, does the SRST measure speaking ability similar to the range of skills assessed by other speaking tests? Although each speaking test is designed to meet purposes and situations in local contexts with varying operationalization of constructs and task characteristics, it is reasonable to assume that some aspects and constructs are commonly measured across tests and thus correlate between them. We will examine this question in this paper.

Relationships between new tests and fairly well-established tests (external criteria) have been examined as part of validation processes (e.g., Messick, 1996). When tests are thought to assess similar abilities, moderate or high correlations are considered concurrent evidence for validity concerning new tests. One such investigation was done by Bernstein, Van Moere, and Cheng (2010), who reported strong relationships between the Versant™ tests and oral interview tests in Spanish, Dutch, Arabic, and English as a second language (L2). For instance, among 130 L2 English learners in Iran, correlations were high between the Versant English and the International English Language Testing System (IELTS; r = .77), the Versant English and the Test of English as a Foreign Language Internet-based Test (TOEFL iBT; r = .75), and the IELTS and the TOEFL iBT (r = .73). They argued that these strong relationships between the Versant™ tests and other tests suggest high validity of interpretation based on the Versant™ test scores. Concurrent validation often attracts criticism: The external criteria tests are often presumed to have perfect validity, which of course they do not (e.g., Bachman, 1990). However, this method is considered appropriate when the result is regarded as just one example of validity evidence, and when this method is used together with other methods for accumulating validity evidence.

Comprehensive test validation requires the demonstration of theoretical and empirical evidence (e.g., Messick, 1996). According to Chapelle, Enright, and Jamieson (2008), while theoretical evidence is usually obtained by describing (a) the importance of the target domain and the relevance and representativeness of the tasks, empirical evidence can be collected by investigating the following: (b1) the appropriateness of the rating scale and

the statistical properties of tasks and ratings; (b2) the reliability of the test and usefulness of the test specifications; (b3) consistency between actual test-taking processes and test developers' intentions, agreement between the difficulty order and the predicted order of the test tasks, and reasonable correlations between the target test and other tests assessing similar or different constructs; (b4) sound relationships between the target test and indicators of ability or real-life performance that the test scores are intended to predict (e.g., speaking ability or real-life performance); and (b5) the meaningfulness of test scores, the feedback for test users (e.g., teachers and test-takers), and the test's beneficial washback on intended aspects such as learning and teaching. Chapelle et al. demonstrated how evidence regarding (a) to (b5) was gathered for their validity argument for the TOEFL iBT using the argument-based approach to validity. Previous studies of the SRST (Hirai & Koizumi, 2009, 2011; Koizumi & Hirai, 2010) covered (a) to (b2). Further, the current study contributes to (a) by comparing the speaking functions (e.g., expressing opinions) elicited from the SRST versus those from other tests (Versant™ English Test and Standard Speaking Test; hereinafter, Versant and SST) in the discussion of the first research question (RQ1, below). Additionally, it aims to contribute to (b3) through comparison with the Versant and to (b4) through comparison with the SST (see RQ2 to RQ4, below).

Current Study

This study compares the SRST with two other speaking tests, the Versant and SST (see the Method section for details) to examine the validity of score-based interpretation of the SRST. The Versant and SST were selected because they have been thoroughly investigated with multiple sources of validity evidence reported (e.g., Nakano, 2002; Pearson Education, 2008) and are now used fairly widely in Japan. Moreover, the SRST, Versant, and SST seem to measure similar aspects of speaking ability.

This study investigates the similarities and differences in the three tests using multiple analytical methods. It should enable test users to grasp the strengths and weaknesses of each speaking test and to select appropriate speaking tests that are relevant to their purpose or situation. Four research questions are addressed:

RQ1: How do the speaking functions elicited by the SRST compare with those elicited by the Versant and the SST?

RQ2: To what extent are SRST scores related to Versant and SST scores?

RQ3: Are there differences in score distributions of the three tests between two groups: beginner and intermediate level learners combined versus higher proficiency level learners?

RQ4: What factors contribute to differences in the scores of the three tests?

For RQ2, given similar and differing test constructs and formats, correlations between the three speaking tests are expected to be moderate.

Method

Participants

Participants were 64 L2 learners of English, consisting of 40 undergraduates and 24 postgraduates from three universities in Japan (28 males, 36 females). Most were between 18 and 24 years of age and were majoring in English, art, culture, or physical education. The participants included 62 students from Japan and one each from China and France.

To investigate RQ2, the 64 test-takers were divided into either (a) a beginner and an intermediate or (b) a higher proficiency-level group, having regard to their majors, educational qualifications, and self-reported proficiency scores. All students who were specializing in English at graduate school (n = 23) and undergraduate students who had self-reported scores of 860 or higher on the TOEIC® (n = 3) were assigned to (b). Hence, 26 test-takers were assigned to (b). The remaining 38 students were assigned to (a). Although stratifying students on the basis of scores achieved in the same test would have been better, we failed to obtain such scores for all test-takers. When we compared students who reported their TOEIC® scores, we found that group (a) had a higher mean (M = 826.92; Median = 890.00; SD = 142.27; n = 13) than group (b) (M = 595.70; Median = 570.00; SD = 124.77; n = 10), Mann-Whitney U = 14.00, Z = -3.16, exact p < .001, effect size r = -.66 (a large effect size).

Tests Used

The procedure for administering the SRST has been described. We employed three stories of similar difficulty (Flesch-Kincaid Grade Level of 4.1 to 4.6), with the story lengths being short to relatively long (94 to 153 words). They were derived from past administrations of the EIKEN (Test in Practical English Proficiency), Grades 3 and 4, and were relatively easy to comprehend. One short story was used for practice, and the other two were

used for the analysis (see Appendix A for a longer story). This number was considered acceptable because Hirai and Koizumi (2011) showed that two stories can sustain sufficient reliability. The order of the two main stories was counterbalanced. It took approximately 22 minutes for SRST test-takers to finish the retelling of the three texts.

The Versant aims to assess "facility in spoken English—that is, the ability to understand spoken English on everyday topics and to respond appropriately at a native-like conversational pace in intelligible English" (Pearson Education, 2008, p. 7). Although this test is intended to assess "the core skills that are building blocks of speaking proficiency" (Bernstein et al., 2010, p. 371), which include both listening and speaking, we focus on speaking assessment and refer to it as a type of speaking test. The Versant is a semi-direct test conducted over the telephone or computer for about 15 minutes and consists of six tasks, including answering questions and retelling stories (see Table 1). Test-takers listen, then start speaking with virtually no planning time. Their utterances are recorded and scored by a fully automated scoring system in which human rating patterns are incorporated, and test results become accessible within minutes. An overall score is derived along with subscores for Sentence Mastery, Vocabulary, Fluency, and Pronunciation. These are reported on a scale of 20 to 80.

Table 1. Structure of the Versant

Part	Task	Number of items
A	Reading: Read a sentence aloud	8
B	Repeat: Listen to a sentence and repeat it	16
C	Short Answer Questions: Listen to a general knowledge question and answer it	24
D	Sentence Builds: Listen to three groups of phrases and reorder them into an understandable sentence	10
E	Story Retelling: Listen to a story and retell it	3
F*	Open Questions: Listen to a question eliciting an opinion and state an answer	2

*Not included in the final score but the sound files are accessible to test users.

The SST is a modified version of the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI), adjusted for Japanese learners of English (ACTFL–ALC Press, 1996; ALC Press, 2010).

Compared with the OPI, the SST has tasks that are more structured and more intermediate levels (three levels for Novice, five levels for Intermediate, and one level for Advanced) based on the ACTFL Proficiency Guidelines (ALC Press, 2010). The aim of the SST is to assess "functional speaking ability" (oral proficiency) and elicit face-to-face interaction that "simulates authentic conversation" (ACTFL–ALC Press, 1996, pp. 1-3) between a certified interviewer and an interviewee. The SST is "adaptive to the perceived level of the examinee as well as his/her personal and professional interests" (ALC Press, 2010). In other words, during simulated conversation that fits the test-taker's interests, the interviewer informally evaluates the test-taker's level based on his/her responses and selects tasks appropriate to the level. For this purpose, the interviewer finds a level at which the test-taker can consistently perform well and identifies "a ceiling of proficiency through prompts designed to elicit from the candidate speech acts at a level higher than s/he has thus far demonstrated" (ACTFL–ALC Press, 1996, p. 7).

According to ALC Press, (2010), the SST is completed in 10 to 15 minutes and comprises five stages: Warm-up questions, Single picture, Role-play with the interviewer, Picture sequences, and Wind-down questions. The recorded conversation is scored by at least two qualified raters. Test-takers receive a holistic score from Levels 1 to 9 with feedback in terms of Global tasks/functions, Communication with interviewer, Text type, Accuracy, Pronunciation, and Comment from interviewer.

Procedures

The participants took three speaking tests (SRST, Versant, and SST) on the same day or on separate days within 2 weeks. We counterbalanced the order of the three tests as far as the schedule allowed, but this was only possible for some candidates.

The Versant was administered using a telephone or computer, depending on the university settings. Before the test, each examinee received an individualized test sheet with test instructions, examples of tasks, and sentences for reading aloud. They had time to read through the sheet and prepare for the test, practicing tasks alone or using examples on the website. For the SST, test-takers met individually in a room with an interviewer and took the test.

Analyses

For RQ1, a checklist of speaking functions was made on the basis of O'Sullivan, Weir, and Saville (2002). Since their final checklist (Appendix 3 of their paper) contained only language functions elicited by the UCLES (University of Cambridge Local Examinations Syndicate) Main Suite examinations, other functions that O'Sullivan et al. omitted but which are observed in real life were included and used for the present analysis. The first author judged whether each function is elicited by all the tests. As a second rater conducting an independent evaluation, the second author, who is well acquainted with the SRST, judged the SRST functions, while another rater familiar with the Versant and SST judged their functions. Inter-rater reliability of all the judgments was high (Agreement ratio = .88; Kendall's tau-b = .82, $p < .001$). After examining the reasons for divergent points, we decided to use our more reasoned judgments as final ones. The open question section of the Versant, whose performance is not scored, was not included in the judgments.

For RQ2, the scores of two stories of the SRST were averaged. To calculate the inter-rater reliability of the SRST, four raters (two English teachers at secondary and tertiary levels and two TESOL graduate students) underwent a one-hour rating training in the use of the EBB scale and benchmark performances. After the training, 16 test-takers out of the 64 (25%) were evaluated by two raters independently. Pearson product-moment correlation coefficients between the two raters' ratings were found to be relatively high ($r = .81$ for CE; $r = .78$ for G&V; $r = .74$ for Pronunciation). Thus, the rest of the test-takers' responses were scored by only one of the raters, and these scores were used for analysis (our limited resources prevented us from asking two raters to evaluate all the students). The reliability of the three SRST criteria (e.g., CE) was found to be high ($\alpha = .84$). Then, all the three rating criteria were summed to produce the total SRST scores. Two examinees failed to complete one of the two stories in the SRST; their scores were imputed using the mean values of all the rest of the examinees.

A sequential multiple regression analysis was conducted using SPSS (Version 12.0.1) to examine the proportion of variance in the SRST scores (dependent variable) explained by the other test scores (independent variables). The sample size of 64 was not very large for multivariate analyses, but it was considered acceptable to use multiple regression analysis since it exceeded the minimum sample size required ($n = 63$) when a study has two independent variables with a medium effect size of R^2, a power of .80, and an alpha level of .05 (Green, 1991).

With regard to RQ3, we made histograms of the two proficiency groups and compared the score distributions to scrutinize each test's capability of discriminating between group members.

For RQ4, we took the following three steps. First, we converted the raw scores of the three tests to standard scores to enable direct score comparisons. Second, in order to examine the test performances of participants who showed large discrepancies among the three test scores, we calculated three types of subtractions in the standard scores by calculating (a) the SRST scores minus the Versant scores, (b) the SRST scores minus the SST scores, and (c) the Versant scores minus the SST scores. While a large number of cases (94%, 180/[64*3]) had similar standard scores (within the value of -1.50 to 1.50), 12 cases (6%; n = 10) showed different standard scores (with the absolute value being more than 1.50). Lastly, we transcribed the utterances of these 12 cases when the recordings were accessible. The performances that were accessible and analyzed were those of the two stories of the SRST, the Story Retelling Task of the Versant, and the overall interview of the SST. Interpretable differences are presented in the Results section.

Results

Comparison of Functions Elicited Using the Checklist

Table 2 shows that although the tests intend to assess aspects of speaking ability, overlapping functions were limited. For example, "describing" and "paraphrasing" were the only functions constantly (as indicated by O) or mostly (as indicated by Δ) elicited by the tests; "elaborating" was elicited by the SRST and mostly by the SST but not by the Versant. No functions were elicited only by the SRST.

Table 2. Functions Elicited by the Three Tests

	Descriptions	SRST	Versant	SST
Informational functions				
Providing personal information	Give information on present circumstances, past experiences, and future plans	Δ	X	O
Expressing opinions	Express opinions	O	X	Δ

	Descriptions	SRST	Versant	SST
Elaborating	Elaborate on, or modify an opinion	O	X	Δ
Justifying opinions	Express reasons for assertions s/he had made	Δ	X	O
Comparing	Compare things/people/events	X	X	Δ
Complaining	Complain about something	X	X	Δ
Speculating	Speculate	X	X	Δ
Staging	Separate out or interpret the parts of an issue	X	X	X
Making excuses	Make excuses	X	X	Δ
Describing	Describe a sequence of events and a scene	O	O	O
Paraphrasing	Paraphrase something	O	O	Δ
Summarizing	Summarize what s/she has said	X	X	X
Suggesting	Suggest a particular idea	X	X	Δ
Expressing preferences	Express preferences	Δ	X	O
Interactional functions				
Agreeing	Agree with an assertion made by another speaker (apart from 'yeah' or nonverbal)	X	X	Δ
Disagreeing	Disagree with what another speaker says (apart from 'no' or nonverbal)	X	X	Δ
Justifying/ Providing support	Offer justification or support for a comment made by another speaker	X	X	X
Modifying	Modify arguments or comments made by other speaker or by the test-taker in response to another speaker	X	X	X
Asking for opinions	Ask for opinions	X	X	X
Persuading	Attempt to persuade another person	X	X	Δ

	Descriptions	SRST	Versant	SST
Asking for information	Ask for information	X	X	Δ
Conversational repair	Repair breakdowns in interaction	X	X	Δ
Negotiating meaning	E.g., check understanding and ask for clarification when an utterance is misheard or misinterpreted	X	X	Δ
Managing interaction functions				
Initiating	Start any interactions	X	X	Δ
Changing	Take the opportunity to change the topic	X	X	X
Reciprocating	Share the responsibility for developing the interaction	X	X	X
Deciding	Come to a decision	X	X	Δ
Terminating	Decide when the discussion should stop	X	X	X

Note. Functions and expressions used here are based on O'Sullivan et al. (2002). O = intended to elicit from all test-takers; Δ = intended to elicit from most test-takers or test-takers at higher levels; X = intended to elicit from a very limited number of or no test-takers.

Correlation and Multiple Regression Analyses

Table 3 shows that the three test scores were normally distributed. Correlations were moderate between the SRST and Versant (r = .64, p < .01) and between the SRST and SST (r = .66, p < .01). A high correlation between the Versant and SST (r = .79, p < .01) accords with Bernstein et al. (2010), who documented strong correlations between the Versant™ tests and various oral interviews in four languages (e.g., r = .77 to .92).

Next, all the assumptions for the multiple regression analysis were checked and confirmed to have been met. Table 4 shows that 43% (adjusted R^2) of the SRST scores were predicted by the SST scores alone and an additional 3% by the Versant scores. Similarly, 41% of the SRST scores were explained by the Versant scores solely, with an additional 5% explained by the SST scores. Overall, the SRST scores were substantially (46%) predicted by the scores of the other two tests. In other words, it was found that there is a general tendency that a candidate scoring high on the Versant and SST is

also likely to have a high SRST score. The finding that 43% of the SRST scores were predicted by the SST scores also means that 43% of the SST scores were predicted by the SRST scores, which suggests that the SRST scores can predict 43% of the SST scores.

Table 3. Descriptive Statistics of the Three Test Scores (*N* = 64)

	Mean	*SD*	Minimum	Maximum	Skewness	Kurtosis	Possible score range
SRST	9.83	2.47	3.00	14.50	-0.81	1.17	3-15
Versant	39.94	10.70	22.00	75.00	1.01	1.52	20-80
SST	4.59	1.61	2.00	9.00	0.88	0.69	1-9

Table 4. Regression Analyses for Predicting the SRST Scores

Variable	R^2	Adjusted R^2	SEE	*F* Change	Change *p*	*F*	*p*
SST only	.44	.43	1.87	47.87[a]	<.01	47.87[a]	<.01
SST + Versant	.48	.46	1.81	4.81[b]	.03	27.81[c]	<.01
Versant only	.41	.41	1.90	43.95[a]	<.01	43.95[a]	<.01
Versant + SST	.48	.46	1.81	7.24[b]	.01	27.81[c]	<.01

Note. SEE = Standard error of estimate. [a] (1, 62), [b] (1, 61), [c] (2, 61).

The finding that more than 40% of the SRST score variance can be explained by the other two tests suggests that the speaking ability assessed by the SRST may be similar to that assessed by the Versant and SST. Additionally, it indicates that more than half of the SRST variance is unexplained, suggesting each test measures distinctive test constructs. While the measurement error (e.g., test-takers' different conditions while taking the tests) could explain this difference, other factors, in addition to the abovementioned elicited different functions, are explored below.

Differences in Score Distributions

Table 5 shows score distributions for each test. For example, on the SRST, three test-takers received scores ranging from 3.00 to 3.99, whereas one had scores between 4.00 and 5.99. On the Versant, seven test-takers obtained scores ranging from 20 to 29. On the SST, three test-takers received Level 2 scores. Patterns become conspicuous in Figure 2, wherein the same

information as in Table 5 is displayed. For the beginner and intermediate level group, the Versant and SST had similar distributions, in which most students obtained lower scores. By contrast, the SRST overall dispersed students of the same ability group across the whole score range. For the higher proficiency group, scores of the Versant and SST were generally distributed within the whole score range, but the SRST scores were not observed in the lower end of the score range (i.e., 3.00 to 5.99). Overall, while the SRST seems to identify differences in the speaking abilities of students up to the intermediate level, the Versant and SST seem to better discriminate between highly proficient students.

Table 5. Score Range and Number of Students

SRST	3	4-5	6-7	8-9	10-11	12-13	14-15	
Beginner/ Intermediate[a]	3	1	4	15	11	4	0	
Higher[b]	0	0	1	5	10	8	2	
All[c]	3	1	5	20	21	12	2	
Versant	20-	30-	40-	50-	60-	70-		
Beginner/ Intermediate[a]	7	22	9	0	0	0		
Higher[b]	1	5	9	9	0	2		
All[c]	8	27	18	9	0	2		
SST	2	3	4	5	6	7	8	9
Beginner/ Intermediate[a]	3	13	12	8	2	0	0	0
Higher[b]	0	1	5	8	5	2	3	2
All[c]	3	14	17	16	7	2	3	2

Notes. SRST: 3 = 3.00-3.99; 4-5 = 4.00-5.99. Versant: 20- = 20-29; 70- = 70-80. SST: 2 = Level 2. [a]*n* = 38; [b]*n* = 26; [c]*N* = 64.

Some may wonder how the SRST could cover the broad score range for the beginner and intermediate learners, considering that it elicits utterances from a limited range of tasks. We believe that the SRST has this capability for two reasons. First, the SRST tends to elicit relatively long utterances, even from lower proficiency students, by presenting model language through the reading passage that can be used for production and by providing them with time to plan their speech; thus their speaking abilities can be well examined

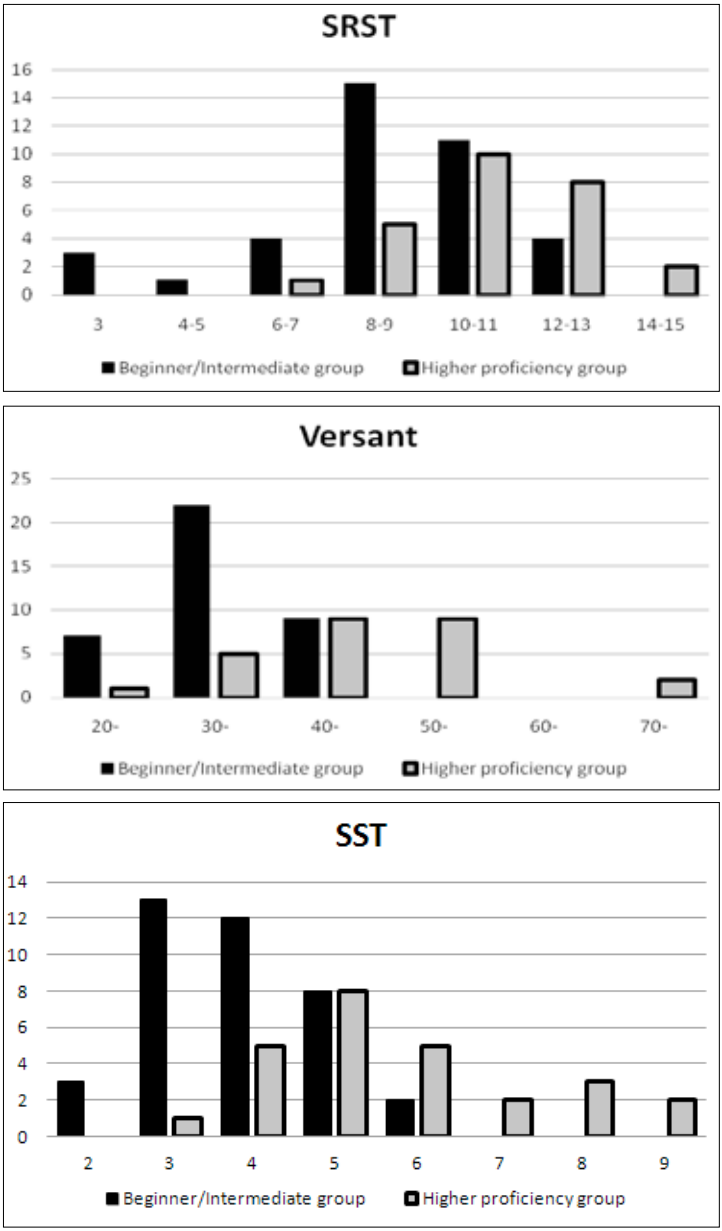


Figure 2. Score Distribution of the Three Tests

and discriminated across score levels. Second, the EBB rating scale for the SRST was empirically developed on the basis of utterances from novice and intermediate level learners, which maximized the discriminatory power of the SRST's EBB scale for such learners. However, this characteristic may vary depending on the difficulty of the stories that test-takers retell. The current study used relatively easy texts. Future studies should examine whether the use of more difficult texts leads to different discriminatory patterns.

Analysis of Transcribed Spoken Data

While the score distributions differ depending on the test-takers' proficiency levels, comparing the transcribed utterances of the three tests indicated two main factors that lead to score differences: (a) tasks and speaking functions and (b) aspects focused on while rating. The first factor, (a), was observed between the SRST and Versant, between the SRST and SST, and between the Versant and SST. First, the SRST and Versant seem to differ in terms of tasks, especially planning time. Examinees can take more time to plan future utterances in the SRST than in the Versant. The SRST does not have a specific time for planning, but candidates can think after they have finished reading a text. In contrast, the Versant gives virtually no planning time and requires quick responses. In one instance, a female student had a higher SRST score (Standardized = 0.44; Raw = 11.00; CE = 4.00; G&V = 4.00; Pronunciation = 3.00) than the Versant score (Standardized = -1.16; Raw = 31.00). She managed to explain the stories (with many pauses) within the specified time during the SRST; however, in the Story Retelling Task of the Versant, she could not finish the stories within the specified time. Her performances seemed to diverge substantially depending on the time allowed.

Between the SRST and SST, there was a case in which a task difference led to dissimilar performances, which resulted in different scores. The SRST has a story retelling and opinion stating task, whereas the SST calls for more varied and complex types of functions in response to an interviewer's prompts, especially for intermediate and advanced level learners. One female examinee achieved a higher SRST score (Standardized = 0.64; Raw = 11.50; CE = 4.00; G&V = 3.50; Pronunciation = 4.00) than SST score (Standardized = -0.99; Level 3, Novice High). She did fairly well in describing the stories she read during the SRST but produced fragmentary utterances and exhibited much difficulty in executing simple tasks, such as explaining her wish about an overseas tour to a tourist agency (in the role-play) during the SST. Given her successful performance in describing picture sequences during the SST, she seems to have the ability to express simple ideas in English when

the specific content to talk about is supplied; however, she is unlikely to have the ability to produce language while simultaneously considering the content.

Another task difference was noted between the Versant and the SST. The SST requires test-takers to use complex functions such as explaining details and giving solutions to problems by employing strategic skills, whereas the Versant's tasks are simpler. One female test-taker had a higher score on the Versant (Standardized = 2.76; Raw = 75.00) than the SST (Standardized = 0.87; Level 6, Intermediate Mid). She succeeded in retelling the gist in a story retelling task on the Versant. In contrast, during the SST, she could not execute her task in a role-play; she failed to convey to a shop clerk her request to exchange a product she had bought. In other SST tasks, she tended to stop when expressing details and complicated concepts. These divergent performances between the Versant and SST seem to suggest that she likely lacks strategic skills to manage and maintain interactions and the ability to describe details, which are elicited in the SST and led to a comparatively lower score on that test.

A second factor that seems to contribute to diverging scores is aspects focused on while rating, which was observed only between the SRST and SST. One female student who obtained a lower SRST score (Standardized = -0.74; Raw = 8.00; CE = 4.00, G&V = 1.00, Pronunciation = 3.00) than SST score (Standardized = 0.87; Level 6, Intermediate Mid) received a low score on grammar and vocabulary on the SRST because her performance contained relatively numerous minor errors. Minor errors were also obvious during the SST; however, her talk was intelligible and convincing with high fluency on the SST, which resulted in a higher SST score. The SRST uses three criteria; when one of the three yields a lower score, the total derived by adding the three scores results in a lower score. On the contrary, in the holistic rating system the SST employs, if test-takers make themselves understood very effectively and achieve the set tasks, they can gain higher scores despite some minor errors in utterances in terms of grammar and pronunciation. The SST holistic scale is weighted more towards communicatively effective performance than minor errors, while the SRST EBB scale gives equal weight to each of communicative efficiency, grammar and vocabulary, and pronunciation. Thus, the SST holistic scale may be able to compensate for minor errors with impressive holistic performance.

These two key factors, (a) tasks and speaking functions and (b) aspects focused on while rating, seem to lead to differences in the evaluation of test performance and the resulting scores, which could invite varied decisions

based on the scores. However, recall that as much as 46% of the SRST score variance was shared by the other two tests (see the *Correlation and Multiple Regression Analyses* subsection). Therefore, we can conclude that the three tests tend to produce close results overall, with some variation caused by the aforementioned factors.

Discussion and Conclusion

The SRST is a semi-direct speaking test devised for classroom use and for measuring the ability to produce extended spoken monologues. This study investigated the relationships between the SRST, Versant, and SST to probe the validity of score-based interpretation of the SRST. RQ1 was "How do functions elicited by the SRST compare with those elicited by the Versant and SST?" We found that when we considered both O and Δ, the SRST elicited more functions (k = 7) than the Versant (k = 2), but fewer than the SST (k = 20). Few functions were shared: The SRST shared two functions with the Versant and seven functions with the SST, while the Versant shared two functions with the SST. The SST was found to elicit more functions by providing several tasks (e.g., picture sequences, role-play) and chances for test-takers to interact with an interviewer. Although the functions elicited by the SRST were limited compared with the functions elicited by the SST, we intended to limit the functions and tasks to focus on areas that Japanese learners of English find difficult and to increase the practicality for administration. Similarly, the Versant elicited a limited number of functions, which corresponds with the developers' intentions to elicit "core skills that are building blocks of speaking proficiency" (Bernstein et al., 2010, p. 371) without using many real-life functions.

RQ2 asked to what degree the SRST scores are associated with the Versant and SST scores. The results showed that correlations of the SRST with the Versant and SST were moderate (r = .64 to .66), that a substantial proportion of the SRST score variance (46%) was predicted by the other two tests, and that the SST alone explained the score variance as much as the Versant did (43% vs. 41%). Such relationships were expected on the basis of intended test constructs and formats, and were empirically supported by the moderate to strong correlations; hence, it is concluded that the SRST likely assesses some of the "facility in spoken English," measured by the Versant and some of the "functional speaking ability," tested by the SST. Moreover, the result—43% of the SST scores was explained by the SRST—seems to show that the construct that the SRST measures is related to real-life interactive communication, since the SST aims to simulate natural conversation.

RQ3 examined differences in score distributions of the three tests between the beginner and intermediate-level learner group combined and the higher level learner group. Figure 2 showed that differences existed in score distributions of the three tests between the two proficiency groups: The SRST scores of the beginner and intermediate-level learners ranged widely, whereas their Versant and SST scores clustered at the lower end of the score range. Conversely, the Versant and SST differentiated higher level learners within the possible score range better than the SRST. These results suggest that the SRST can better differentiate speaking abilities in beginner and intermediate-level students, whereas the Versant and SST can better discriminate such abilities in students of higher proficiency.

RQ4 aimed to identify factors contributing to score differences in the three tests by analyzing the transcripts of test-takers' utterances. The analysis indicated that (a) tasks and speaking functions and (b) aspects emphasized while rating could cause score differences. As for task differences, the SRST allows story retelling and opinion statement after the possibility of some planning time, whereas the Versant asks test-takers to perform several simple tasks immediately after they are provided. The SST elicits fluent use of interactive functions using various tasks such as talking about familiar topics, stating opinions, negotiating, and elaborating on details and complex matters, depending on test-takers' levels. With respect to differences in scoring systems, the SST rating focuses more on fluent and effective communication than on errors that do not impede understanding, whereas the SRST concentrates equally on communicative efficiency and accuracy aspects.

Three implications are discussed. First, this study contributed to the accumulation of validity evidence for the SRST and demonstrated one instance of the validation process by providing multiple new strains of empirical evidence derived through comparison with the other tests. This study and previous ones (Hirai & Koizumi, 2009, 2011; Koizumi & Hirai, 2010) covered most critical analyses in the validation framework, as delineated in the introductory section (i.e., regarding the [a] to [b4] aspects). However, investigation into the meaningfulness of the test scores, the feedback to test users, and the beneficial washback of the test on intended aspects such as learning and teaching (i.e., [b5] in the framework above) remains to be done. The impact of the SRST on learning speaking skills, especially when used as a formative and summative assessment tool in language classrooms, should specifically be inspected.

The second implication is that this study explained one difference between the three tests (i.e., score distributions between the beginner/intermediate

group and the higher proficiency group) and two factors differentiating the scores (i.e., tasks and speaking functions, and rating method). This information may provide a useful basis for selecting one of the three speaking tests as appropriate for a given assessment purpose and testing situation. Although qualified interviewers are needed along with test budgets, the SST typically elicits interactive and complex functions by assigning test-takers various tasks that fit their proficiency levels, while focusing on effective communication. The Versant, while requiring monetary and equipment resources (telephones or computers), measures natural-paced listening along with the ability to react promptly. Further, the SST and Versant tend to discriminate between learners of a higher proficiency group. On the other hand, the SRST requires teachers or peers to evaluate performances using the EBB rating scale, and it uses a limited range of tasks (i.e., story retelling and opinion stating) and elicits a limited number of functions (i.e., providing personal information, expressing opinions, elaborating, justifying opinions, describing, paraphrasing, and expressing preferences). However, it has three chief advantages, particularly when used as a classroom test. First, it is free. Second, teachers can incorporate the test into classroom activities by using passages that students have already learned. Third, it is likely to discriminate between speaking performances, particularly in students at beginner and intermediate levels. Thus, the SRST might be capable of effectively discriminating between students who have achieved speaking goals in class from those who have not, and of demonstrating students' short-term speaking development. These three advantages indicate that, for the purpose of formative assessment, teachers can provide feedback regarding aspects that students have been taught (e.g., pronunciation) during the lessons, conduct remedial activities, and test the same aspects again to scrutinize the improvement, when their resources allow them to do so. These classroom uses of the SRST might encourage speaking activities comprising extended monologues and enhance the speaking ability of L2 learners, although this needs to be empirically tested. It may be worthwhile for teachers to use the SRST for their class, considering its advantages and its shared aspects with the Versant and SST.

A third implication is that, although the SRST is primarily constructed as a test for classroom settings, the relatively large proportion of variance shared by the Versant and SST might indicate that retelling tasks are useful in other settings. The TOEFL iBT already has such speaking integrated tasks, in which examinees read or listen to texts and speak based on the information provided (Chapelle et al., 2008). The Versant also utilizes a story

retelling task based on a listening stimulus. While there are variations in text types (academic and nonacademic) and specific activities elicited (oral summary vs. retelling as much as possible; with or without adding opinions), generally retelling tasks could be useful with novices, intermediates, and advanced learners.

Finally, researchers should modify two aspects of the current procedures to obtain stronger evidence for validity in future studies. First, it is necessary to counterbalance the order in which the three speaking tests are taken to offset possible order effects. Second, the criterion for dividing test-takers into two proficiency groups should be improved. This study stratified students, primarily using information about students' majors and educational qualifications, with minor adjustments to accommodate their self-reported proficiency scores. However, separating the two groups on the basis of the same proficiency test scores would better clarify actual proficiency levels and provide more meaningful interpretations of the results. More rigidly generated evidence for validity would strengthen the SRST validity argument and enhance the usefulness of the SRST in the classroom context.

Acknowledgement

This research was partially supported by the Grant-in-Aid for Scientific Research (KAKENHI, C) [grant number 19520477]. We would like to thank the following for their contributions to our paper: Knowledge Technologies, Pearson, for letting us use the Versant and its sound files; ALC Press, for providing sound files of the SST; Yujia Zhou and Emiko Kaneko for their assistance in test administration and analysis; and Yo In'nami for his critical comments regarding our draft.

Rie Koizumi is an Assistant Professor at Tokiwa University. *Akiyo Hirai* is a Professor at the University of Tsukuba. They are interested in validating speaking tests.

References

ACTFL–ALC Press. (1996) *Standard Speaking Test manual.* Tokyo: Author.

ALC Press. (2008). *TSST: About the test format and assessment.* Retrieved from http://tsst.alc.co.jp/tsst/e/assessment.html

ALC Press. (2010). *The Standard Speaking Test (SST).* Retrieved from http://www.alc.co.jp/edusys/sst/english.html

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27,* 355-377. doi:10.1177/0265532210364404

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language™.* New York: Routledge.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing.* Cambridge: Cambridge University Press.

Educational Testing Service. (2011). *TOEIC® Speaking and Writing: About the tests: Test content.* Retrieved from http://www.ets.org/toeic/speaking_writing/about/content/

Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education*. Cambridge: Cambridge University Press.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26,* 499-510.

Hirai, A., & Koizumi, R. (2009). Development of a practical speaking test with a positive impact on learning using a story retelling technique. *Language Assessment Quarterly, 6,* 151-167. doi:10.1080/15434300902801925

Hirai, A., & Koizumi, R. (2011). *Validation of empirically-derived rating scales for the Story Retelling Speaking Test*. Unpublished manuscript.

Koizumi, R., & Hirai, A. (2010). Exploring the quality of the Story Retelling Speaking Test: Roles of story length, comprehension questions, keywords, and opinions. *ARELE (Annual Review of English Language Education in Japan), 21,* 211-220. Retrieved from http://ci.nii.ac.jp/els/110008512411.pdf?id=ART0009707205&type=pdf&lang=jp&host=cinii&order_no=&ppv_type=0&lang_sw=&no=1322450319&cp=

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13,* 241-256. doi:10.1177/026553229601300302

Nakano, M. (2002). Standard Speaking Test (SST) to TOEIC, TOEFL, EIKEN tono kaiki bunseki [Regression analysis of Standard Speaking Test (SST), TOEIC, TOEFL, and the EIKEN]. *Research report from Institute of Oral Communication, Waseda University* (pp. 23-50). Tokyo: Institute of Oral Communication, Waseda University. Retrieved from http://www.alc.co.jp/edusys/sst/pdf/article3.pdf

National Institute for Educational Policy Research of Japan. (2007). *The investigation on the special project 'English speaking.'* Retrieved from http://www.nier.go.jp/kaihatsu/tokutei_eigo/05002051033004000.pdf

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19,* 33-56. doi:10.1191/0265532202lt219oa

Pearson Education. (2008). *Versant™ English Test: Test description and validation summary.* Retrieved from http://www.versanttest.com/technology/VersantEnglishTestValidation.pdf

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49,* 3-12. doi:10.1093/elt/49.1.3

Zen mondai & kaito 2001 nendo dai 3 kai kentei: Ichiji shiken [All test items and answers at the third EIKEN in the academic year of 2001: First stage]. (2002). *STEP the Latest on English*, *24*, 1-74.

Appendix A

Instructions and a Story Sample

Read the story silently within two minutes. 2分間で次の文章を黙読しなさい。

Story 2

Last year, Bob and his sister Jean went to Florida for their summer vacation with their parents. They visited Florida for one week. The weather was very nice. So everyone was really happy.

On the first day, Bob and his family went to the beach. It was beautiful. The sand was white and the water was very clean. Bob and Jean swam for over three hours. After that, they played volleyball with some other children on the beach. Their parents watched them and smiled.

After playing volleyball, Bob and Jean felt tired. They sat down on the sand next to their parents and drank cold coconut water. Their father said, "It's getting late. Let's go back to the hotel." But Bob and Jean didn't want to leave the beach. Jean asked, "Can we come back again tomorrow?" Their mother said, "Of course we can." Bob and Jean were very happy to hear that.

(Zen mondai & kaito 2001 nendo dai 3 kai kentei: Ichiji shiken, 2002, p. 59; Copyright 2002 by the Society for Testing English Proficiency, Printed with permission)

After the signal, read each question aloud and answer them in English.
1問ずつ合図があってから、質問を読み上げて、英語で答えなさい。

Q1: What did Bob and his family do on the first day?

Q2: How long did Bob and Jean swim?
Q3: Why were Bob and Jean happy?

----------<Next Page>----------

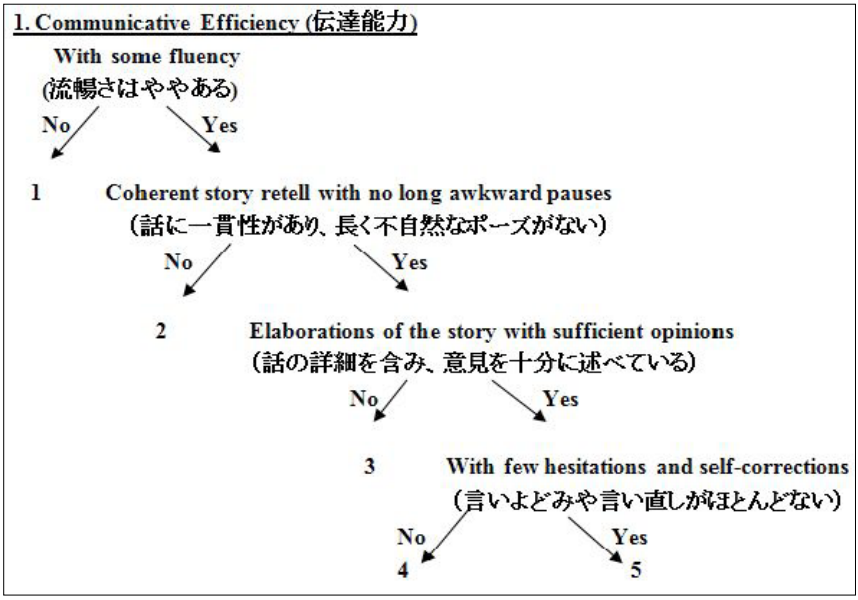
Retell as much of the story as you can in English in two and half minutes. You can look at the keywords while you are retelling. At the end of your retelling, be sure to include your opinions about the story. You will hear a signal 30 seconds before closing.

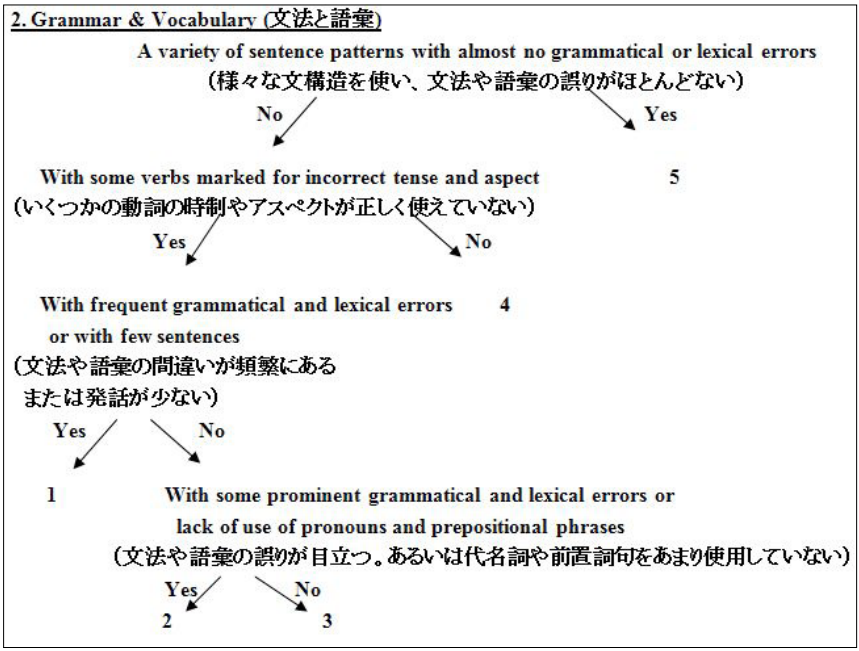
今読んだ内容をできるだけ詳しく、2分30秒間英語で話しなさい。話しながら、キーワードを見てもかまいません。読んだ内容を話し終えたら、必ず、その内容についての感想や意見も英語で述べなさい。終了30秒前にチャイムがなりますので、感想を始める目安にできます。

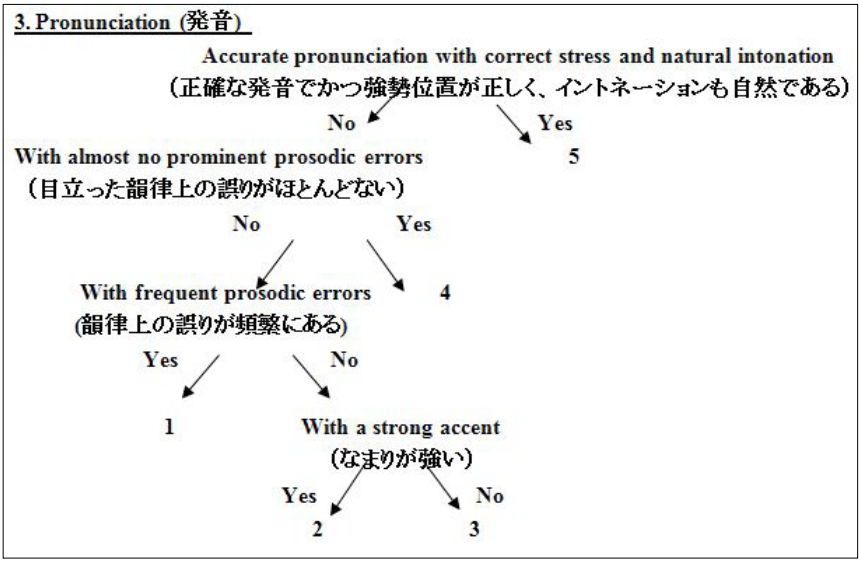
Keywords: Bob, Jean, Florida, beach

Appendix B

EBB rating scale for the SRST




2. Grammar & Vocabulary (文法と語彙)
A variety of sentence patterns with almost no grammatical or lexical errors
(様々な文構造を使い、文法や語彙の誤りがほとんどない)
No
Yes
With some verbs marked for incorrect tense and aspect
5
(いくつかの動詞の時制やアスペクトが正しく使えていない)
Yes
No
With frequent grammatical and lexical errors
4
or with few sentences
(文法や語彙の間違いが頻繁にある
または発話が少ない)
Yes
No
1
With some prominent grammatical and lexical errors or
lack of use of pronouns and prepositional phrases
(文法や語彙の誤りが目立つ。あるいは代名詞や前置詞句をあまり使用していない)
Yes
No
2
3


3. Pronunciation (発音)
Accurate pronunciation with correct stress and natural intonation
(正確な発音でかつ強勢位置が正しく、イントネーションも自然である)
No
Yes
With almost no prominent prosodic errors
5
(目立った韻律上の誤りがほとんどない)
No
Yes
With frequent prosodic errors
4
(韻律上の誤りが頻繁にある)
Yes
No
1
With a strong accent
(なまりが強い)
Yes
No
2
3

Use of *I* in Essays by Japanese EFL Learners

Sayo Natsukari
Rikkyo University

A problematic issue in Japanese EFL learners' academic writing is the overuse of the personal pronoun *I*. The use of personal pronouns is particularly important in academic writing because it determines the writer's perspective and attitude toward the readership. This study investigates the extent to which Japanese EFL learners' use of the first-person singular *I* in essays is different from the norms of native speakers. By using subcorpora of the International Corpus of Learner English and the Louvain Corpus of Native English Essays, the study compares the use of *I* in argumentative essays. The results indicate that Japanese EFL learners use *I* in a similar way to American students, but they overuse *I* in essays: Almost all essays by Japanese learners contain *I* and the number of *I*'s are excessive. The analysis also uncovers an excessive use of the phrase *I think* in their essays.

日本人英語学習者のアカデミックライティングにおける問題点のひとつとして主語のIの多用がある。本稿は、約20万語から成るICLE書き言葉コーパスを利用し、議論形式のアカデミックエッセーにおいて、日本人英語学習者の主語Iをどのように使用しているのか調査し、これをアメリカとイギリスの大学生が書いたアカデミックエッセーから構成されたLOCNESS書き言葉コーパスのデータと比較した。結果、日本人学生の人称代名詞のIの使用方法はアメリカ人学生のそれに類似していたが、全体的な過剰使用が確認された。特に、ほぼすべての学生がIを使用し、その量も過剰であった。I thinkというフレーズも母語話者の書き言葉に比べて過剰であった。本稿は、この過剰使用の問題点を議論する。

JALT Journal, Vol. 34, No. 1, May 2012

The grammatical rules for personal pronouns are simple, but those governing their use in writing are rather complex. The choices between personal pronouns (first, second, and third person, as well as singular and plural) create different relationships between writers and readers. First- and second-person pronouns imply an interpersonal relationship between the author and audience, whereas third-person pronouns generate an impersonal context. Thus, the use of these personal pronouns differs according to the purpose of writing and the intended audience.

The use of personal pronouns is complex in academic writing. According to Pennycook (1994), pronouns are "complex and political words, always raising difficult issues of who is being represented" (p. 173). In addition, Kuo (1999) points out that personal pronouns name a self, selves, and others; therefore, their use determines the writer's position and attitude in academic society.

In particular, the use of the first-person pronoun *I* in academic prose is controversial and has often been investigated. Conventionally, *I* is considered less common in academic prose because it sounds conversational, informal, egocentric, and less objective (Biber, Johansson, Leech, Conrad, & Finnegan, 1999; Chafe, 1982; Korhonen & Kusch, 1989; Kuo, 1999; Smith, 1986). However, some studies on personal pronouns in written discourse have discussed views that *I* is sometimes acceptable depending on the occasion and the manner in which it is used (Smith, 1986; Tang & John, 1999), as well as on preferences of the individual and academic discipline (Coniam, 2004; Hyland, 2001; Petch-Tyson, 1998; Smith, 1986).

Accordingly, textbooks and guidebooks for academic writing treat the use of the first-person pronoun *I* with caution. Basically, the first-person singular is not common in academic writing (Fowler & Aaron, 2010; Johns, 1997; Langan, 2000), but some textbooks point out that it has been becoming more common in recent years in certain appropriate parts of an essay (Cooley & Lewkowicz, 2003; Fowler & Aaron, 2010). Nevertheless, the use of first-person singular in academic writing still greatly depends on the discipline (Bergmann, 2010; Fowler & Aaron, 2010), and thus students are recommended to ask university supervisors whether or not the use of *I* is accepted in their course of study (Creme & Lea, 2008).

Academic essays by nonnative speakers have been reported as lacking balance and having greater subjectivity (Hinkel, 1999). Some corpus-based research on English academic essays written by Finnish and Swedish (Herriman & Aronsson, 2009; Petch-Tyson, 1998) as well as Japanese (Akahori, 2007; Ishikawa, 2008) EFL learners has found an overuse of first-person

pronouns. Some researchers have also noted that *I think* is excessively used in essays by EFL learners from Japan (Ishikawa, 2008; Oi, 1999a), from France (Aijmer, 2002), and from several other European countries (Ringbom, 1998). When EFL learners exhibit problems in their use of personal pronouns in essays, investigating how their personal pronoun use differs from the norms of native speakers can help teachers instruct their students on how to improve their academic writing.

First-Person Pronouns and Essays by Japanese EFL Learners

One of the issues that Japanese EFL learners face in academic writing is the appropriate use of first-person pronoun singular *I*. Some instructors and researchers report that Japanese EFL learners use *I* very frequently. In their instructional textbook for Japanese EFL learners on how to write English academic essays, Kamimura and Oi (2004) note that Japanese EFL learners use *I* too frequently, which can make their writing lack objectivity and generality.

There are some small-scale reports on the use of *I* in academic writing by Japanese EFL learners. Oi (1999a) compared an academic composition by an American university student with those of two Japanese university students, all of which were on the same topic. She pointed out that, unlike the American student, the Japanese students frequently used *I* and wrote openly about their personal lives. Suganuma (2004) conducted a classroom investigation into the writing of 44 Japanese university students. Her analysis of their persuasive essays reveals that all students regularly used first-person pronouns, including *I*, *my*, *me*, and *myself*, mainly in two ways: expressing opinions and describing experiences.

An investigation into the use of personal pronouns (Akahori, 2007) also discovered the excessive use of *I* in argumentative essays by Japanese EFL learners. The essays were written by Japanese undergraduate and postgraduate students, and were collected as part of learner corpora in the International Corpus of Learner English (ICLE). The first version of the ICLE was published in 2002, but a Japanese subcorpus was not included because the essays then written by Japanese EFL learners were not fully argumentative. Akahori compares Japanese students' writings for ICLE with those of American and British university students in the Louvain Corpus of Native English Essays (LOCNESS). Her analysis indicates that one of the reasons Japanese EFL learners' writings lack argumentativeness may be their subjective perspectives, which is seen in excessive use of *I* (Akahori, 2007, p. 5).

Another corpus-based study (Ishikawa, 2008) investigated frequent words and phrases in opinion essays using the Corpus of English Essays Written by Japanese University Students (CEEJUS), finding that the personal pronouns *I* and *we*, as well as the phrase *I think*, are very frequently used.

This study attempts a more detailed and wider scale investigation into the use of *I* in argumentative essays by Japanese EFL learners, drawing on ICLE version 2 (ICLEv2; Granger, Dagneaux, Meunier, & Paguot, 2009), which includes written English data produced by Japanese undergraduate and postgraduate students as a Japanese EFL subcorpus. Building on the investigations into the frequency of *I* that have been conducted in previous reports and studies, this study will explore how the use of *I* in argumentative essays by Japanese EFL learners is different from native speaker norms.

Method

Data

The data of Japanese EFL learners' written language used in this study comes from ICLEv2 (Granger, et al., 2009). ICLEv2 includes argumentative essays of a higher English proficiency level written by EFL learners of 16 different first-language backgrounds, who come from all over the world. The Japanese subcorpus consists of 366 argumentative essays written by Japanese undergraduate and postgraduate students, all of whom were native speakers of Japanese. The total number of words is 198,241.

The two sets of data of native speakers' written language come from LOCNESS, which comprises British pupil A-level essays, British university student essays, and American university student essays (Granger & De Cock, n.d.). Data of argumentative essays written by university students were extracted from LOCNESS for comparison with Japanese university student argumentative essays. There were 33 argumentative essays by British university students, amounting to 19,019 words, and 175 by American students, amounting to 149,574 words. Each essay in ICLEv2 and LOCNESS was written by a different student. Table 1 shows the description of subcorpora: ICLEv2 Japanese, LOCNESS British, and LOCNESS American.

Essays in ICLEv2 and LOCNESS cover a wide range of topics. Some of the most frequent topics in ICLEv2 are English education, technology, human rights, and social issues such as environment, crime, and gender. Similarly, essays by American students in LOCNESS include a variety of issues such as euthanasia, capital punishment, and animal testing. The 33 argumentative

essays by British students in LOCNESS, on the other hand, are on a single topic: "A single Europe: A loss of sovereignty for Britain."

Table 1. Description of the Subcorpora

Subcorpus	Number of essays	Number of words	Number of running words with AntConc	Mean number of words per essay
ICLEv2 Japanese	366	198,241	202,099	552
LOCNESS British	33	19,019	19,042	577
LOCNESS American	175	149,574	150,544	860

Analysis

This study investigates the frequency of *I* and how *I* is used (i.e., functions of *I*) in argumentative essays. The frequency of *I* was analyzed using a concordance tool called AntConc (3.2.1w; Anthony, 2007). Three sets of comparison were made in order to examine to what extent the frequency of *I* in academic essays varied among three corpus groups: the occurrences of *I* per 1000 words, the number of essays including *I*, and the number of essays that overuse *I*. The datum line of overuse was determined by cluster analysis using SPSS version 15, where data of native speakers were divided into two groups according to the number of occurrences of *I* per 1000 words. In this study, the number of essays overusing *I* is equivalent to the number of writers who overuse *I* since, in ICLEv2 and LOCNESS databases, one author writes only one essay.

Although the frequency was determined automatically by the concordance tool, the functions of *I* were analyzed manually. In this study, *I* is classified into four categories to analyze its occurrences in context. (Note: Errors in learners' writings have not been corrected in ICLEv2.)

(1) I *for Personal Matters*

This feature includes *I* used for descriptions of the author's personal matters, such as personal identity and experience. Writers often use *I* to write about their personal status, ability, and situation.

I am a 21 year old male. (ICLE-US-IND-0018.1)

Instances of *I* for expressing the writer's personal experiences and actions in the present or past are also included in this category as personal matters.

I entered university. (ICLE-JP-TM-0007.1)

I send her an email what I want to say. (ICLE-JP-SWU-0005.4)

Writers write about their feelings, hopes, and knowledge in their life.

I know a boy. (ICLE-JP-WA-0005.1)

I hope my dream come true. (ICLE-JP-SWU-0014.4)

These cases are included in this category, but the writer's feelings and hopes regarding the issue in the argumentative essays are classified into category (2), which is when *I* is used to write about personal opinions and feelings about an issue in question.

(2) I *for Opinions*

The second category includes the writer's use of *I* to argue an opinion or indicate feelings on an issue.

I think the effect is far more damaging ... (ICLE-BR-SUR-0015.3)

(3) I *for Organizing*

In this category, *I* is used in the context of the organizational structure of the essay or to guide readers. The use of *I* explicitly shows the organization and structure of the essay. It also draws attention to important points and guides readers from one point to another throughout the essay.

I am going to write about good points of both ... (ICLE-JP-RI-0002.1)

(4) I *Used in Conversation*

The final category of *I* is seen in direct conversational quotations.

"Yes, I am." (ICLE-JP-SWU-0019.4)

Results

The Frequency

Frequency analysis shows that *I* occurs about five to ten times more often in Japanese EFL learners' essays than in native speaker essays. When the frequency is measured as the number of occurrences per 1000 words (Table 2), it is found that Japanese EFL learners use *I* 21.15 times per 1000 words, whereas British and American students use *I* 2.83 and 4.44 times per 1000 words, respectively.

The analysis of the number of essays including *I* also shows that most argumentative essays written by Japanese EFL learners include *I*. Out of the 366 essays they authored, 95.90% (351) include *I*. On the other hand, 54.55% and 58.86% of British and American student essays, respectively, contain *I*.

Table 2. Instances of *I* in Essays by Japanese, British, and American Students

	ICLEv2 Japanese	LOCNESS British	LOCNESS US
Number of essays	366	33	175
Instances of *I*	4275	54	668
Maximum number of *I* in an essay	73	12	32
Minimum number of *I* in an essay	0	0	0
Standard deviation	11.21	2.53	6.02
Instances of *I* (per 1000 words)	21.15	2.83	4.44
Number of essays including *I*	351 (95.90%)	18 (55.55%)	103 (58.86%)
Mean of *I* per essay	11.68	1.64	3.82

Looking at the individual data, I found that some students use *I* more excessively than others. On the basis of cluster analysis, both the British and the American data were divided into two groups, depending on whether or not the data included *I* seven times or more per 1000 words. Since the average number of occurrences of *I* per 1000 words in British and American data were 2.83 and 4.44 respectively, it seemed reasonable to choose seven or more occurrences of *I* per 1000 words as the criterion for overuse.

Such excessive users of *I* are a minority of native speakers, but a majority of Japanese. In the American data, 39 essays out of 175 (22.29%) showed an overuse of *I*. In the British data, five essays (15.15%) showed overuse. Most instances of *I* occurred in the work of a limited number of excessive users. In the American data, 75.30% of *I* use is in 22.29% of essays, and in the British data, 59.26% of *I* use is in 15.15% of essays.

On the other hand, in the Japanese data, 284 of 366 essays (77.60%) showed an overuse of *I*. The number of students who did not overuse *I* was 82 (22.40%), including those who did not even use *I* at all (15). Approximately 96% of Japanese learners used *I* in their essays, and nearly 80% overused it.

Different Uses in Context

The functions of *I* in essays was analyzed to discover how *I* is used (Table 3). Approximately 60% of *I* was used for writing about the writers' personal matters in the American and Japanese essays, but this was rarely seen in the British essays (11.11%). Japanese and American essays were also similar in the use of *I* to convey the author's opinions (32.0% and 27.99%, respectively) and in conversational quotations (1.43% and 1.95%, respectively). On the other hand, the British students mainly used *I* when they wrote about their opinions: 77.78% of instances of *I* in essays by British students were for this purpose. Direct conversational quotations were absent from the British students' essays examined here. The single similarity among Japanese, American, and British students is the usage of *I* for organizing the essays. In each group, approximately 10% of instances of *I* were used in this manner.

Table 3. The Purposes of Using *I*

	Personal matters	Opinions	Organization	In conversation	Total
Japanese ICLE v2 (Total 366 / Excessive users 284, 77.60%)					
Frequency	2531	1368	314	62	4275
(Excessive users)	(2509)	(1277)	(290)	(60)	(4136)
%	59.20%	32.0%	7.35%	1.45%	100%
(Excessive users)	(58.69%)	(29.87%)	(6.78%)	(1.40%)	(96.74%)
Per 1000 words	12.52	6.77	1.55	0.31	21.15
(Excessive users)	(16.00)	(8.15)	(1.85)	(0.38)	(26.38)

	Personal matters	Opinions	Organization	In conversation	Total
British LOCNESS (Total 33 / Excessive users 5, 15.15%)					
Frequency	6	42	6	0	54
(Excessive users)	(6)	(23)	(3)	(0)	(32)
%	11.11%	77.78%	11.11%	0	100%
(Excessive users)	(11.11%)	(42.59%)	(5.56%)	(0)	(59.26%)
Per 1000 words	0.31	2.21	0.31	0	2.83
(Excessive users)	(2.08)	(7.97)	(1.04)	(0)	(11.09)
American LOCNESS (Total 175 / Excessive users 39, 22.29%)					
Frequency	397	187	71	13	668
(Excessive users)	(334)	(114)	(47)	(8)	(503)
%	59.43%	27.99%	10.63%	1.95%	100%
(Excessive users)	(50%)	(17.06%)	(7.04%)	(1.20%)	(75.30%)
Per 1000 words	2.64	1.24	0.47	0.09	4.44
(Excessive users)	(9.96)	(3.40)	(1.40)	(0.23)	(14.99)

Excessive Users of I

Although the use of *I* by Japanese and American students is similar in the percentages of how *I* is used in context, the biggest difference is that the excessive use of *I* was seen only in a limited number of essays in the native speakers' writing, while it was seen in most of the Japanese data.

The most frequent use of *I* in American essays was for writing about personal matters, for which it was used in 44.57% of essays, but most of these instances (84.13%) appeared in the writing of excessive users, who make up 22.29% of the American data. Likewise, the use of *I* by excessive writers outnumbered the use of *I* by all other writers in the other categories: *I* for expressing opinions, showing the organization of the essay, and in conversation (Table 3).

Similarly in British essays, *I* was used primarily by the five excessive writers, who made up only 15.15%. *I* for personal matters was used only in essays by these excessive writers. In particular, the overuse of *I* in British data was limited to a single use—to express opinions—and 56.76% of *I* for this use was seen in the excessive writers' essays.

On the other hand, 77.60% of the Japanese students were excessive users of *I*. Overall, more than 90% of *I* for each type of use was accounted for by the excessive users.

Excessive Use of I Think

Further investigation uncovered the excessive use of *I think* in essays by the Japanese EFL learners. As Table 4 displays, *think* was the most frequent word to collocate with *I* in Japanese essays. *I* co-occurs with *think* 4.30 times per 1000 words, which account for 20% of the total occurrences of *I*. The Japanese learners over-depended on using *I think*. In essays by native speakers, on the other hand, *think* was a frequent word associated with *I*, but other verbs that made up phrases used to show the author's "stance" (Biber et al., 1999), such as *feel* and *believe*, were also used (see Table 4).

Table 4. Words That Collocate With *I* (Immediately Preceding or Following)

Rank	ICLEv2 Japanese		LOCNESS British		LOCNESS American	
	Frequency*	Collocate	Frequency*	Collocate	Frequency*	Collocate
1	867 (4.30)	think	10 (0.53)	feel	62 (0.41)	have
2	214 (1.06)	don't	6 (0.32)	think	42 (0.28)	feel
3	206 (1.02)	want	5 (0.26)	would	41 (0.27)	think
4	195 (0.96)	can	5 (0.26)	believe	37 (0.25)	was
5	182 (0.90)	was	4 (0.21)	am	32 (0.21)	would
6	179 (0.89)	am	3 (0.16)	have	30 (0.20)	'm
7	171 (0.85)	have	2 (0.01)	use	27 (0.18)	am
8	102 (0.50)	will	2 (0.01)	can	24 (0.16)	know
9	82 (0.41)	'm	1 (0.05)	wouldn't	22 (0.15)	believe
10	75 (0.37)	would	1 (0.05)	therefore	18 (0.12)	had

* per 1000 words

Discussion

The results of this wider scale corpus-based study confirm what previous studies claim: that *I*, as well as *I think*, is overused in essays by Japanese EFL learners (Akahori, 2007; Ishikawa, 2008; Oi, 1999a). In addition, the use of *I* in context is found to be similar to Suganuma's (2004) results, which discovered that the first-person pronouns are mainly used for expressing opinions (30%) and for stating the writer's personal experience (70%). The comparison with native speakers' essays in this study revealed that *I* appears occasionally in essays by native speakers, and that the number of

essays in which *I* is overused is limited, whereas *I* is overused in most of the essays by the Japanese EFL learners.

Use of I *and Overuse of* I

Although native speakers of English at tertiary education institutions do not use *I* nearly as frequently as Japanese EFL learners do, this study reveals that some of them occasionally use it in their academic essays. The first-person pronoun *I* appeared in 54.55% and 58.89% of essays by the British and American students, respectively. As textbooks and guidebooks for academic writing report, it is common to see the occasional use of *I* in students' essays.

There are, however, some differences between writing by British students and American students. Biber (1987) finds more cases of *I* and *you* pronouns in American academic prose than British academic prose, and this study also found more cases of *I* in American students' essays than British students' essays. One possible explanation given by Biber is that *I* and *you* pronouns are factors indicating an interactive and colloquial style, and American writing has a "greater use (or tolerance) of informal, colloquial, and interactional features" (p. 113). In addition to the colloquial language use in American writing, this study suggests that American and British students use the personal pronoun *I* for different purposes. Both British and American students use *I* to express opinions, but American students in particular write about their personal matters in essays. Personal experience is considered to be acceptable in "academic composition in Anglo-American educational environments" (Hinkel, 1999, p. 91), but there might be differences in the use of *I* even among native speakers with different cultural and educational backgrounds, as a study by Connor & Lauer (1988) suggests.

Although the use of *I* is for the most part acceptable, the overuse of *I* could spoil a piece of academic writing, regardless of the writer's native language, from two points of view: (a) academic tone and (b) subjectivity. An impersonal and formal tone is preferred in professional, academic writing; overuse of *I* may disconcert readers by making the text seem conversational, interactive, and less formal. The use of *I* indicates the writer's involvement with the audience (Chafe, 1982), which is also referred to in several other ways, such as presence of writer (Hyland, 2001; Smith, 1986) and writer/reader visibility (Petch-Tyson, 1998); the frequent use of *I* is a common feature of spoken language and is therefore less formal (Smith, 1986). Investigations into academic writing confirm this: *I* is not used in scientific articles (Kuo, 1999). In addition, Korhonen and Kusch (1989) and Kuo find that in

many argumentative texts, the plural *we* is more dominant as compared to the singular *I*.

Although this study does not aim to investigate the rhetoric in writing, the quality of rhetoric can be considered partly from the use of *I*. Essays with frequent *I*, with occasional *I*, and without *I* show some differences in their argumentation. Most essays with excessive use of *I* argue an issue from personal experience throughout the essay; by contrast, in writing where *I* is used only occasionally, personal experience is only one among various resources for argument. Writing about personal matters is a feature of personal writing rather than academic writing (Creme & Lea, 2008). Writers who frequently use *I* for expressing opinions typically discuss an issue only from their personal connection to it. Sometimes the whole essay and at other times only part of the essay is heavily argued from a subjective point of view. In essays with little or no *I*, writers usually mention several points of view, for example, for and against or pros and cons. They express their opinions indirectly but clearly by advocating or criticizing other ideas. Thus, such essays often include a detailed analysis of the issue.

Japanese Learners

It is a problem that most Japanese university students overuse *I* in their academic writing. In other words, Japanese learners of English write academic essays in conversational, less formal, interactive, personal, and subjective tones, and in so doing may risk putting off readers from other backgrounds. Several factors are believed to account for their overuse of *I* in academic essays, including the following: (a) Japanese language and Japanese writing culture, (b) lack of writing skills as language learners, and (c) textbooks written for EFL learners.

Investigations into the rhetoric patterns of writing in English by Japanese learners have revealed that they are influenced by the rhetoric patterns appropriate to Japanese writing (Hinds, 1983; Oi, 1999b). The overuse of *I*, resulting in too much presence of the writer in an essay, could also be influenced by the Japanese language and Japanese writing culture. It is sometimes argued that the Japanese language is a predicative-oriented language, and it is syntactically very different from Western languages (Morita, 2002). Japanese discourse is always expressed from a speaker's point of view and the speaker is not always present in the discourse, unlike English discourse, in which the speaker is always identified. Oi (1999a) and Kamimura and Oi (2001) discuss the fact that English is a subject-predicate type language,

while the Japanese language is typologically a topic-comment type: Comments on the topic in Japanese discourse are provided from the speaker's point of view. Therefore, if Japanese students try to translate what they want to say in Japanese into English, the easiest subject for them to use is *I*.

Since Japanese discourse reflects the speaker's perspective, the typical and traditional composition written in Japanese is full of subjectivity. In school, the typical composition is based on the student's personal experience or on a topic close to the student, and consists of the details of an event and the student's personal opinions or feelings concerning it. Whereas the main aim of composition writing in American elementary schools is the development of writing skills and techniques such as essay organization, Japanese elementary schools aim at nurturing personality through expressing personal experiences and feelings (Watanabe, 2004). Such traditional composition writing in Japan influences the English writing skills of Japanese learners (Oi, 1999a; Suganuma, 2004).

Japanese learners' excessive use of the subjective phrase *I think* to express opinions is also likely influenced by the Japanese language (*I think* = *omou*). The main definitions of *I think* and *omou* are similar to each other. In Japanese *omou* is used to show opinions as well as uncertainty (Moriyama, 1992). While *omou* can also be translated as *one may think* or *it is thought*, virtually all writers, using *I*, apply it by default as the subject of *think*. Similarly, in English, *I think* is "used when you are saying that you believe something is true, although you are not sure" (*Longman Dictionary of Contemporary English*, 2003, p. 2014). Such similarities must make it easy for Japanese EFL learners to say *I think*.

Another reason for the excessive use of *I think* can be explained from the rhetoric pattern of English written by Japanese students. According to rhetoric studies, Japanese learners tend not to assertively express their opinions. Kamimura and Oi (1998) argue that Japanese students excessively use *I think* as a softening device, while American students occasionally use *I think* as an emphatic device. In their study, 80% of Japanese learners used *I think* before stating their opinions. Oi (1999b) discovered that Japanese learners use indecisive argumentation in their essays. They write their thinking process in essays. They use *I think* to show opinions along with their thinking process, resulting in indecisive argumentation. The overuse of *I think* reflects a direct translation from *omou* and the indecisive manner of writing in the Japanese language and culture.

The overuse of *I* and a subjective tone are not limited to writing by Japanese EFL learners; they also appear in the writing of EFL learners from

various other first-language backgrounds. Hvitfeldt (1992) finds that the writing of many Malay students is personalized and includes descriptions of their personal lives. She explains that personalization is commonly seen in writing from oral-oriented cultures, yet she also argues that the same feature is found in writing by "students who have not yet made the shift from the oral discourse style to the more literate discourse style" (p. 38), regardless of their native language. The shift is likewise depicted in Ivanič and Camps (2001); a Mexican student shifted from her preference for using first person to the use of an impersonal style of writing accepted by academics. Furthermore, a series of studies that investigated the learners' written language using ICLE corpus data found a higher frequency of language use that shows writer or reader visibility (Petch-Tyson, 1998) and a higher frequency of the subjective phrase *I think* in writings by European EFL learners (Aijmer, 2002; Herriman & Aronsson, 2009; Ringbom, 1998), which leads the researchers to conclude that this reflects the trait of a conversational tone in learners' writing. According to Biber et al. (1999) and Biber and Reppen (1998), the phrase, *I think* as well as *I think that* are typically spoken phrases, and they are rarely seen in academic prose. Finally, the overuse of *I* can be explained from the simple structural repertoire of EFL learners. As seen in Table 4, the *be*-verb often co-occurs with *I* in Japanese learners' writing, which reflects the over-statement of personal matters in their compositions. This *be*-copula as a feature of simple syntax is commonly seen in writings by nonnative speakers (Hinkel, 2003). The overuse of *I* and *I think* observed in EFL learners shows that they are in the process of language learning.

Finally, the influence of textbooks and instructions for writing at the tertiary level in Japan should be considered. There are many textbooks that provide ample writing exercises and model essays where the writer's life and opinions are at the center of the description. Textbooks emphasizing "the cognitive process of writing" often focus on the students' personal experiences and interests as a topic (Spack, 1988). In addition, personal-expressivist and learner-centered views (Johns, 1997), whose focus is to develop students' fluency and confidence in writing, encourage students to write about personal experiences and thoughts. Such practices are common in textbooks used in EFL classrooms, but it is questioned whether such personal writing exercises actually help students to develop the academic writing skills required outside language lessons (Johns, 1997; Spack, 1988). A background full of personal writing could affect the writer's choice of the first person in his or her academic writing.

Conclusion

This study investigates the details of the overuse of *I* in argumentative essays written by Japanese EFL learners using the Japanese subcorpus in ICLEv2. Most Japanese learners overuse *I* in English academic essays. They particularly overuse *I* to write about personal matters and to express their opinions. In order to show their opinions, the phrase *I think* is excessively used. The findings imply that the overuse of *I* in essays by Japanese EFL learners is influenced by their linguistic and cultural background as well as lack of academic writing skills. Although the use of personal pronouns is a small factor in academic writing, teaching how to use them will help improve the EFL learners' academic writing in terms of objectivity, maintaining an impersonal perspective, and formality.

One of the limitations in this study is that the sizes of the subcorpora are not equal; in particular, the amount of corpus data from British university students is smaller than the amounts from Japanese and American students. Therefore, diversity between American and British students is not conclusive. Nonetheless, the investigation into the three subcorpora provides informative implications for teaching.

The findings should be taken into account when English academic writing instruction is given to Japanese EFL learners, especially at the university level. First, teachers should help students develop their overall academic writing skills. For example, students may know the grammatical rules for personal pronouns but not necessarily be aware of the appropriateness of their usage or the perspectives that they connote. Both linguistic correctness and appropriateness for the context are important in academic writing. In addition, teachers can raise students' awareness of differences between English academic writing and Japanese techniques that they have learned before entering university. Additionally, in teaching rhetoric or argumentation in English at the micro-level, teachers should help students learn how to use a wider variety of linguistic expressions to enable them to write opinions and show organization of an essay both with and without *I*. Last but not least, textbooks for EFL courses and language input from reading should be selected more carefully.

Note

The original version of this paper was presented at JALT2010 in Nagoya, Aichi (Natsukari, 2010).

Sayo Natsukari is currently an adjunct lecturer at Rikkyo University.

References

Akahori, N. (2007). The excessive use of the first-person pronoun "I" in English compositions by Japanese students. *Nihonjin eigo gakushusha no hanashikotoba kakikotoba no corpus sakusei to sono goyoron teki taishobunseki* [Building learner corpora of Japanese learners of English and the contrastive analyses]. (Report for Grant-in-Aid for Scientific Research (B) No. 15320059). Tokyo: Brainsnetwork.

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Grenger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 55-76). Amsterdam: John Benjamins.

Anthony, L. (2007). *AntConc* (Version 3.2.1w) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Bergmann, L. S. (2010). *Academic research and writing*. Boston: Longman.

Biber, D. (1987). A textual comparison of British and American writing. *American Speech, 62*(2): 99-119.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finnegan, E. (1999). *Longman grammar of spoken and written English*. London: Pearson Longman.

Biber, D. & Reppen, R. (1998). Comparing native and learner perspectives on English grammar: A study of complement clauses. In S. Granger (Ed.), *Learner English on computer* (pp. 145-158). New York: Longman.

Chafe, W. L. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 35-53). Norwood, NJ: Ablex.

Coniam, D. (2004). Concordancing yourself: A personal exploration of academic writing. *Language Awareness*, *13*, 49-55.

Connor, U., & Lauer, L. (1988). Cross-cultural variation in persuasive student writing. In A. C. Purves (Ed.), *Writing across languages and cultures* (pp. 138-159). Newbury Park, CA: Sage.

Cooley, L., & Lewkowicz, J. (2003). *Writing at university: A guide for students* (3rd ed.). Maidenhead, UK: Open University Press.

Creme, P., & Lea, M. R. (2008). *Writing at university* (3rd ed.). New York: McGraw-Hill.

Fowler, H. R., & Aaron, J. E. (2010). *The little, brown handbook* (11th ed.). New York: Pearson Education.

Granger, S., Dagneaux, E., Meunier, F., & Paguot, M. (2009). *International corpus of learner English version 2*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Granger, S., & De Cock, S. (n.d.). *LOCNESS: Louvain corpus of native English essays*. Retrieved from http://www.uclouvain.be/en-cecl-locness.html

Herriman, J., & Aronsson, M. B. (2009). Themes in Swedish advanced learners' writing in English. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 101-120). Amsterdam: John Benjamins.

Hinds, J. (1983). Contrastive rhetoric: Japanese and English. *Text*, *3*, 183-195.

Hinkel, E. (1999). Objectivity and credibility in L1 and L2 academic writing. In E. Hinkel (Ed.), *Culture in second language teaching and learning* (pp. 90-108). Cambridge: Cambridge University Press.

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly*, *37*, 275-301.

Hvitfeldt, C. (1992). Oral orientations in ESL academic writing. *College ESL*, *2*(1), 29-39.

Hyland, K. (2001). Humble servants of the discipline? Self-mention in research articles. *English for Specific Purposes*, *20*, 207-226.

Ishikawa, S. (2008). *Eigo corpus to gengo kyoiku* [English corpus and language education]. Tokyo: Taishukan.

Ivanič, R., & Camps, D. (2001). I am how I sound: Voice as self-representation in L2 writing. *Journal of Second Language Writing*, *10*, 3-33.

Johns, A. M. (1997). *Text, role, and context*. Cambridge: Cambridge University Press.

Kamimura, T., & Oi, K. (1998). Argumentative strategies in American and Japanese English. *World Englishes*, *17*, 307-323.

Kamimura, T., & Oi, K. (2001). The effects of differences in point of view on the story production of Japanese EFL students. *Foreign Language Annals*, *34*, 118-130.

Kamimura, T., & Oi, K. (2004). *Eigo rombun, report no kakikata* [How to write essays and reports in English]. Tokyo: Kenkyusha.

Korhonen, R., & Kusch, M. (1989). The rhetorical function of the first person in philosophical texts—The influence of intellectual style, paradigm and language. In M. Kusch & H. Schröder (Eds.), *Text, interpretation, argumentation* (pp. 61-77). Hamburg: Helmut Buske.

Kuo, C.-H. (1999). The use of personal pronouns: Role relationships in scientific journal articles. *English for Specific Purposes*, *18*, 121-138.

Langan, J. (2000). *College writing skills* (5th ed.). Singapore: McGraw-Hill.

Longman dictionary of contemporary English (4th ed.). (2003). Harlow, UK: Longman.

Morita, Y. (2002). *Nihongo bunpo no hassou* [A conception of Japanese grammar]. Tokyo: Hitsuji Shobo.

Moriyama, T. (1992). Bunmatsu shiko doshi "omou" wo megutte [On mental verb, *omou*]. *Nihongogaku* [Japanese Studies], *11*(9), 105-116.

Natsukari, S. (2010, November). *The use of I of Japanese EFL learners.* Paper presented at the Conference of the Japan Association for Language Teaching, Nagoya, Japan.

Oi, K. (1999a). A note on Japanese students' preference for the first person perspective in writing in English. *Monograph Series*, 37-47.

Oi, K. (1999b). Comparison of argumentative styles: Japanese college students vs. American college students—An analysis using the Toulmin model. *JACET Bulletin*, *30*, 85-102. Retrieved from NII-Electronic Library Service.

Petch-Tyson, S. (1998). Writer/reader visibility in EFL written discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 107-118). New York: Longman.

Pennycook, A. (1994). The politics of pronouns. *ELT Journal*, *48*, 173-178.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger (Ed.), *Learner English on computer* (pp. 41-52). New York: Longman.

Smith, Jr., E. L. (1986). Achieving impact through the interpersonal component. In B. Couture (Ed.), *Functional approaches to writing: Research perspectives* (pp. 108-119). London: F. Pinter.

Spack, R. (1988). Initiating ESL students into the academic discourse community: How far should we go? *TESOL Quarterly*, *22*, 29-51.

Suganuma, A. (2004). Persuasive writing in English by Japanese EFL learners: A study on their frequent use of first person singular pronouns. *The Society of English Studies*, *34*, 21-34. Retrieved from NII-Electronic Library Service.

Tang, R., & John, S. (1999). The 'I' in identity: Exploring writing identity in student academic writing through the first-person pronoun. *English for Specific Purposes*, *18*, S23-S39.

Watanabe, M. (2004). *Nattoku no kozo* [Structure of convincement]. Tokyo: Toyokan.

A Many-Facet Rasch Measurement of Differential Rater Severity/Leniency in Three Types of Assessment

Farahman Farrokhi
University of Tabriz, Iran

Rajab Esfandiari
Imam Khomeini International University, Iran

Edward Schaefer
Ochanomizu University, Japan

Rater effects in performance testing is an area in which much new research is needed (C. M. Myford, personal communication, 23 February, 2010). While previous studies of bias or interaction effect as a component of rater effect have employed experienced teachers as raters (e.g., Schaefer, 2008), the present study uses many-facet Rasch measurement (MFRM) to investigate differential rater effect or rater severity or leniency among three rater types: self-assessor, peer-assessor, and teacher assessor. Essays written in English by 188 Iranian English majors at two state-run universities in Iran were rated both by the students themselves as self-assessors and peer-assessors and by teachers, using a 6-point analytic rating scale. MFRM revealed differing patterns of severity and leniency among the three assessment types. For example, self-assessors and teacher assessors showed the opposite pattern of severity and leniency as compared with peer-assessors when assessing the highest and lowest ability students. This study has implications for the use of peer and self-rating in L2 writing assessment.

JALT Journal, Vol. 34, No. 1, May 2012

評定者効果は今後新たな研究が求められる分野である（C. M. Myford, personal communication, 2010年2月23日）。評定者効果の構成要素の一つであるバイアスや交互作用効果に関する従来の研究は、経験豊富な教員を評定者としたものであるが（Schaefer, 2008）、本研究では多相ラッシュ測定（MFRM）を用い、自己評価、学習者間評価、教員評価の3タイプ間における評定者効果ないし評定者の厳格さ／寛容さを調査した。イランの州立大学2校の188名の英語専攻の学生の書いたエッセイを対象に、自己評価、学習者間評価者として学生たち自身の評価、および教員による6段階分析評定法により評価を行った。MFRMにより、この３つの評価タイプに、厳格さ／寛容さにおいて異なるパターンがあることが明らかになった。例えば、最も能力の高い学生と最も能力の低い学生に対する自己評価と教員評価は、学習者間評価とは反対のパターンの厳格さ／寛容さを示した。以上をふまえL2ライティングの評価における学習者間評価と自己評価使用に関する示唆を与える。

Performance testing involving the use of rating scales has become widespread in the evaluation of second language writing and speaking assessment. With a communicative approach to language teaching, it is felt that this sort of testing gives a fairer reflection of classroom learning and goals than traditional tests. Research into performance testing has focused on student performance, the tests, the scales, and more recently, on raters themselves and what they do when they rate. There has also been interest in the behavior and comparison of different types of raters. One reason for this is that performance testing places an added burden on teachers, since it is more time-consuming than discrete point tests. There has been growing attention paid to the uses of peer-assessment and self-assessment as alternatives or supplements to teacher assessment. If such assessment can be shown to be valid and reliable, it could contribute to lessening the burden on teachers (Fukazawa, 2010).

However, rater judgments do have an element of subjectivity, and this subjectivity has an influence on the reliability and validity of test scores (Eckes, 2009; Lumley, 2005; Schaefer, 2008). Without the implementation of rigorous measurement tools, it is difficult to establish validity or reliability for rater judgments. Two studies that compared rater types but fell somewhat short in this regard are Mahoney (2011) and Yamanishi (2004). In his comparison of teacher and student error evaluations on a dictation quiz, Mahoney found that students tended to evaluate peers' written work more leniently than teachers. Yamanishi's study—in which two groups of raters, high school teachers and university students who were teacher candidates, rated high school students' free compositions—found that while the teachers were consistent in their ratings, the teacher candidates rated somewhat inconsistently. However, these studies relied on raw scores in their analysis and thus lack generalizability. Mahoney acknowledges this when he cau-

tions that in his study comparisons of scores across groups from different classes cannot be made (p. 117).

A promising measurement resource in the investigation of rater behavior in performance testing is many-facet Rasch measurement (MFRM; e.g., Eckes, 2008, 2009; Linacre, 1989/1994), an application of the Rasch model (Rasch, 1960), a logistic latent trait model of probabilities which calibrates the difficulty of test items and the ability of test-takers independently of one another, but places them within a common frame of reference (O'Neill & Lunz, 1996). MFRM expands the basic Rasch model by enabling researchers to add the facet of judge severity to person ability and item difficulty and place them on the same logit (log odds units) scale for comparison. Engelhard (1992) states that MFRM improves the objectivity and fairness of the measurement of writing ability because writing ability may be over- or underestimated through raw scores alone if students of the same ability are rated by raters of differing severity. MFRM adjusts for rater variability and thus provides a more accurate picture of ability. Coniam (2008) observes that "the use of raw scores may substantially disadvantage test takers receiving lower final grades—a situation which in some examination situations may result in failure rather than success on a test" (p. 71). Applying MFRM to data from the Hong Kong Certificate of Education (HKCE) public school examination writing section, Coniam showed that writers of the same ability would get a lower grade if they had a severe rater rather than a lenient rater, thus potentially failing a high-stakes test.

While the majority of published Rasch measurement studies have been conducted in English-speaking countries, Rasch measurement has been attracting increasing attention in Asia, as the 2008 study by Coniam shows. Though it is still not well known in Japan, there have been a number of Rasch studies here as well. Studies that examined peer-assessment of speaking tests with MFRM include Holster (in press) and Fukazawa (2010). Fukazawa used MFRM to investigate the validity of peer-assessment in a Japanese high school setting and concluded that peer-assessment has sufficient validity for assessing speeches in a Japanese high school. However, as in Mahoney (2011), Fukazawa found that student raters rated their peers more leniently than teachers. In a study of peer-assessment of oral presentations given by Japanese university students, Holster used the quality control feature of MFRM known as fit statistics to show that peer raters were highly misfitting, suggesting that they were interpreting the scoring rubric in different ways from teacher raters. He argued that MFRM can be used to provide diagnostic feedback for peer-assessors with idiosyncratic rating patterns, but that

given the high rate of misfit, it is more suited to low-stakes classroom testing rather than high-stakes tests.

Although the studies described thus far examined rater behavior in performance testing, they did not take advantage of another feature of MFRM called bias analysis, which is valuable in studying rater behavior more deeply. As the present study uses bias analysis to investigate rater differences, it is necessary to explain bias analysis in detail.

The bias analysis function of MFRM investigates rater variability in relation to the other facets in the Rasch model. The term bias refers to rater severity or leniency in scoring, and has been defined as "the tendency on the part of raters to consistently provide ratings that are lower or higher than is warranted by student performances" (Engelhard, 1994, p. 98). Wigglesworth (1993) further stated that bias analysis identifies "systematic subpatterns" of behavior occurring from an interaction of a particular rater with particular aspects of the rating situation (p. 309). It can help researchers explore and understand the sources of rater bias, thus contributing to improvements in rater training and rating scale development. In the present study, we use bias analysis to investigate differential rater functioning and rater severity and leniency, in which rater types display favorable or unfavorable inclinations toward either individual students, individual assessment criteria, or items of the rating scale (cf. Du, Wright, & Brown, 1996; Ferne & Rupp, 2007; Knoch, Read, & von Randow, 2007).

Previous Bias Analysis Studies

Bias analysis studies search for unexpected interactions, such as those between rater judgments and test-takers' performance. In one study, Wigglesworth (1993; 1994) looked at rater-item, rater-task, and rater-test type interaction in the speaking portion of the Australian Assessment of Communicative English Skills (*access:*), an English skills test for potential immigrants to Australia. She found significant variation in how raters responded to different test criteria. Some rated grammar more harshly, and others rated it more leniently. Likewise, some raters were stricter on fluency or vocabulary, while others rated these more leniently. Moreover, raters differed from each other in their strictness or leniency towards the different task types. Also in Australia, McNamara (1996), in analyzing the results of the Occupational English Test (OET), found that trained raters were overwhelmingly influenced by candidates' grammatical accuracy, contrary to the communicative spirit of the test, and that the raters themselves were unaware of this. McNamara noted that this study showed the usefulness of MFRM in

revealing underlying patterns in ratings data and fundamental questions of test validity.

Lumley (2005) used MFRM and think-aloud protocols to analyze the writing component of the STEP (Special Test of English Proficiency), another high-stakes test for immigrants to Australia. Initially, MFRM was used to eliminate misfitting raters. Ultimately, four trained raters rated 12 writing samples consisting of two tasks each, for a total of 24 samples, which had been taken for research purposes from a pool of STEP test examination papers. MFRM analyses of these samples found significant differences between raters. Like McNamara (1996), Lumley also found that grammar was the most severely rated category.

In his MFRM study of rater bias patterns in a Japanese EFL setting, Schaefer (2008) employed 40 native English speakers to rate 40 essays written by Japanese university students. Each rater rated all 40 essays, using a 6-point analytic rating scale consisting of five categories (Content, Organization, Style and Quality of Expression, Language Use, Mechanics, and Fluency). The results showed that for 11 of the raters, "if Content and/or Organization were rated severely, then Language Use and/or Mechanics were rated leniently, and vice versa" (p. 465). Schaefer also found that "some raters also rated higher ability writers more severely and lower ability writers more leniently than expected" (p. 465).

Addressing self-assessment, peer-assessment, and teacher assessment, Matsuno (2009) used MFRM with 91 students and four teacher raters to investigate how self- and peer-assessments work in comparison with teacher assessments in actual university writing classes in Japan. He conducted a bias analysis of rater-writer interactions and found that "self-raters tended to assess their own writing more strictly than expected" (p. 91). Moreover, in this study "high-achieving writers did not often rate their peers severely and low-achieving writers did not often rate their peers leniently" (p. 92), but peer-assessors showed "reasoned assessments independent of their own performances" (p. 92). Finally, teacher assessors showed relatively individual bias patterns.

In investigating the phenomenon of rater subjectivity and inconsistency in L2 performance testing, MFRM allows researchers to analyze rater effects at both group and individual levels (Myford & Wolfe, 2004). Bias analysis can identify patterns in ratings unique to individual raters or across raters, and whether these patterns, or combinations of facet interactions, affect the estimation of performance. However, most of these studies have examined individual rater variation. There still do not seem to be many studies that

have investigated the possibility of systematic patterns in rater *type* variation. Eckes (2008) used cluster analysis following an MFRM bias study to identify the existence of rater types, but these types emerged from group scoring profiles, such as Structure Type and Fluency Type. The present study defines type as a preexisting group, that is, self-assessor and peer-assessor (student assessors) and teacher assessor. The only study to our knowledge that has used bias analysis to investigate rater variation in self-assessment, peer-assessment, and teacher assessment (Matsuno, 2009) did not concentrate on rater type patterns at all. Given the paucity of research on systematic bias patterns in rater type, this area warrants further research.

Furthermore, previous bias studies have either dealt with ESL situations as opposed to EFL, or utilized native L1 raters of L2 essays. Negishi (2010) used MFRM to analyze Japanese L1 raters' assessment of the group oral EFL interactions of Japanese secondary and university students (though this was not a bias study), but the other studies reported above all fit this pattern. There is a need for EFL studies that employ nonnative English-speaking raters of EFL essays, since that is the reality of ELT testing in many countries.

The Present Study

In the present study, MFRM was employed to investigate differential rater severity and leniency with nonnative English speaker raters in an EFL situation. We were interested in how three rater types, self-assessor, peer-assessor, and teacher assessor, interact with the assessment criteria or items of the rating scale. Closely related to this, we were also interested in the severity and leniency of rater type toward students. An important implication of this study is the possibility of student peer and self-assessment as an alternative to teacher assessment. If such ratings can be shown to be equivalent, this could be an argument for the use of self- and peer-assessment as a way to reduce teachers' workload (Fukazawa, 2010).

Research Questions

To achieve the goals of the present study, the following research questions are presented.

RQ1: How do self-assessors, peer-assessors, and teacher assessors differ in severity or leniency in relation to items?

RQ2: How do self-assessors, peer-assessors, and teacher assessors differ in severity or leniency in relation to students?

Methodology

Participants

The participants were 194 raters, subdivided into student raters and teacher raters. The student raters were 188 undergraduate Iranian English majors enrolled in advanced EFL writing classes in two prestigious state-run universities in two different cities in Iran, specializing in three fields of study: English Literature, Translation Studies, and English Language Teaching. The student raters were labeled either self-assessors or peer-assessors. The teacher assessors were six Iranian teachers of English.

The student raters ranged in age from 18 to 29, with one over 30, and another with unidentified age. One hundred and thirty-one student raters (69.7%) were female and 57 (30.3%) were male. One hundred and five (55.9%) were native-Farsi speakers, 68 (36.2 %) were native-Turkish speakers, 11 (5.6%) were native-Kurdish speakers and the other four (2.1%) were grouped as "Other." Ninety-five (50.5%) were sophomores, 29 (15.4%) were juniors, and 64 (34.0%) were seniors. Only three (1.6%) of them had the experience of living in an English-speaking country. The number of years they had studied English ranged from 1 to 24, and most of them (61.7%) had studied the English language in language institutes before entering the university.

The teacher assessors were all male, ranging in age from 23 to 36. They came from two language backgrounds: four native-Farsi speakers, and two native-Turkish speakers. None of them had the experience of living in an English-speaking country. They had taught writing courses from 1 to 7 years. Three of them were affiliated with a national university, one of them with a private university, and two were classified as "Other." Each had a degree in English: Three were PhD students in ELT, two had MAs in ELT, and one had a BA in English literature.

The Rating Scale

Generally speaking, there are three types of rating scales in language testing: primary trait, analytic, and holistic (Weigle, 2002). For the purposes of the present study, we chose an analytic rating scale, adapted from Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey's (1981) ESL Composition Profile, but differing from it in many aspects (See Appendix).

To develop our rating scale, we also referred to writing textbooks in the literature because we wanted the scale to reflect the structure of a standard five-paragraph essay, so the following three books were consulted as a guide to composing the scale categories: *Composing With Confidence: Writing Ef-*

fective Paragraphs and Essays (Meyers, 2006), *Refining Composition Skills: Grammar and Rhetoric* (Smalley, Ruetten, & Kozyreve, 2000), and *The Practical Writer with Readings* (Bailey & Powell, 2008). The scale contains 15 items (substance, thesis development, topic relevance, introduction, coherent support, conclusion, logical sequencing, range, word choice, word form, sentence variety, overall grammar, spelling, essay format, and punctuation/capitalization/handwriting).

These 15 items were equally weighted. Although Jacobs et al.'s (1981) five category scales were differentially weighted, it is not clear how those weights were determined (Kondo-Brown, 2002). Hamp-Lyons (1991) recommends using focused holistic scoring when different weights are assigned to different categories in a given context. Schaefer (2008) also observes that different weights predetermine the ranking importance.

There is no consensus in the literature on the optimal number of levels or bands, but for this study, we created a 6-point scoring scale for each item, because "this is the most common number of scale steps in college writing tests, and a large number of steps may provide a degree of step separation difficult to achieve as well as placing too great a cognitive burden on raters, while a lower number may not allow for enough variation among the multifaceted elements of writing skills" (Schaefer, 2008, p. 473).

Data Collection

One hundred and eighty-eight 5-paragraph essays were collected over a year and a half from 188 students. The students came from six classes taught by four instructors, and there were a total of eight weekly meetings. Following the mandatory syllabus set by the Ministry of Sciences, Research, and Technology in Iran, the students were taught punctuation, expression, features of a well-written paragraph, and the principles of a one-paragraph and a five-paragraph essay.

On the midterm exam the week after the last meeting, the students were given 90 minutes to write a five-paragraph essay ranging in length from 500 to 700 words on the following topic, chosen from a list of TOEFL TWE (Test of Written English) topics: *In your opinion, what is the best way to choose a marriage partner? Use specific reasons and examples why you think this approach is the best*. All the students were given the same topic in order to control for topic effect.

Following the data collection, a rating session was held with all the student raters. Before the actual rating, there was a 1-hour training session.

Raters were given an essay rating sheet, one rated essay, and guidelines in Farsi explaining the rating scale in detail. They were told to read the rated essay first without paying attention to the corrections made on the essay. When they finished reading the essay, the researcher conducting the session directed their attention to the corrections made on the essay and the way it was rated on the rating essay sheet. The researcher then explained in detail the rating essay sheet and how the scores had been assigned.

After this, they were given a new essay written by one of the students and told to read the essay and rate it according to the guidelines. They were instructed to closely follow the guidelines, and the researcher monitored the rating process and explained any unclear points.

Following the training session, the actual ratings were held, beginning with self-assessment. The students were given a new rating essay sheet, the guidelines, and their own essays to rate. The researcher advised them to rate as accurately as possible. Following self-assessment, they were given their classmates' essays, with names removed, to rate as peer-assessment. The same rating procedure was repeated. The entire training and rating session took about 2 hours.

The same rating procedure was repeated for teacher assessors. Since it was not possible to arrange for a group meeting, the researcher met with the teachers individually, instructed them how to rate, gave them all 188 essays, and asked them to complete and submit them in one month.

Results

The present study employs a fully crossed design in which all raters rate all essays. The data was analyzed with *Facets 3.68.1*, a software program for MFRM (Linacre, 2011). Three facets were specified for this study: students, rater type, and items.

To answer the first research question (How do self-assessors, peer-assessors, and teacher assessors differ in severity or leniency in relation to items?), a bias analysis between rater type and items was specified in *Facets*. There were 45 bias terms in all. Table 1 shows the cases of significant rater type by items bias.

Table 1. Rater-Type-Items Bias/Interaction Analysis

Rater type	Logit	Items	Logit	Obs. score	Exp. score	Obs-Exp average	Bias size	Model S. E.	*t*-score	Infit MnSq	Outfit MnSq
Self	-.17	2	.08	704	625.2	.55	-.47	.08	-5.56	1.0	1.0
Self	-.17	3	.26	715	584.5	.92	-.77	.09	-8.63	.9	.9
Self	-.17	4	.06	711	628.9	.57	-.50	.09	-5.81	.9	.9
Self	-.17	6	.06	661	628.0	.23	-.18	.08	-2.36	.8	.8
Self	-.17	7	-.32	620	674.3	-.39	.33	.07	4.48	.5	.5
Self	-.17	8	.13	549	619.9	-.49	.32	.07	4.94	.6	.6
Self	-.17	9	.54	460	519.9	-.42	.25	.06	3.90	1.0	1.0
Self	-.17	10	-.22	615	668.3	-.38	.30	.07	4.21	.7	.7
Self	-.17	13	-.59	689	724.1	-.25	.26	.08	3.20	1.1	1.1
Peer	.05	2	.08	617	563.2	.39	-.28	.07	-3.70	.9	.9
Peer	.05	3	.26	621	515.0	.78	-.54	.08	-6.99	1.1	1.1
Peer	.05	7	-.32	594	642.4	-.35	.26	.07	3.71	.6	.6
Peer	.05	10	-.22	589	628.9	-.28	.21	.07	2.95	.7	.7
Peer	.05	15	.16	516	552.1	-.26	.16	.07	2.43	.9	.9
Teacher	.12	2	.08	3902	4034.7	-.13	.08	.02	3.29	1.1	1.1
Teacher	.12	3	.26	3493	3729.5	-.23	.14	.02	5.72	1.4	1.4
Teacher	.12	4	.06	3989	4093.8	-.10	.06	.02	2.60	.9	.9
Teacher	.12	7	-.32	4652	4549.4	.10	-.08	.03	-2.83	1.5	1.5
Teacher	.12	10	-.22	4522	4428.8	.09	-.07	.03	-2.49	.9	.9

Fixed (all = 0) chi-square: 414.6 *df*: 45 significance: .00: $p < .00$

Note. Items: 1 = Substance, 2 = Thesis development, 3 = Topic relevance, 4 = Introduction, 5 = Coherent support, 6 = Conclusion, 7 = Logical sequencing, 8 = Range, 9 = Word choice, 10 = Word form, 11 = Sentence variety, 12 = Overall grammar, 13 = Spelling, 14 = Essay format, 15 = Punctuation.

As is evident in the table, the standard errors (SEs) are low, and the mean square fit statistics are good, with no cases of misfit. Out of 45 bias terms, only 19 were significant, with *t*-scores either greater than +2 or smaller than -2. Eleven of the significant interactions are positive (showing severity), and eight of the significant interactions are negative (showing leniency). Rater type showed significant bias toward only 10 out of 15 items (items 2, 3, 4, 6, 7, 8, 9, 10, 13, and 15). Self-assessor shows nine significant interactions, teacher assessor shows five, and peer-assessor also shows five. All three rater types had slightly more cases of severe bias than lenient (self-assessor: 5 vs. 4; peer-assessor: 3 vs. 2; and teacher assessor: 3 vs. 2). The item displaying

the highest *t*-scores was item 3, with teacher assessor severely biased at 5.72 and self-assessor leniently biased at -8.63. Figure 1 presents a graphical representation of rater differences. It can be seen that the gap is particularly large between self-assessor and teacher assessor on item 3 (Topic relevance).

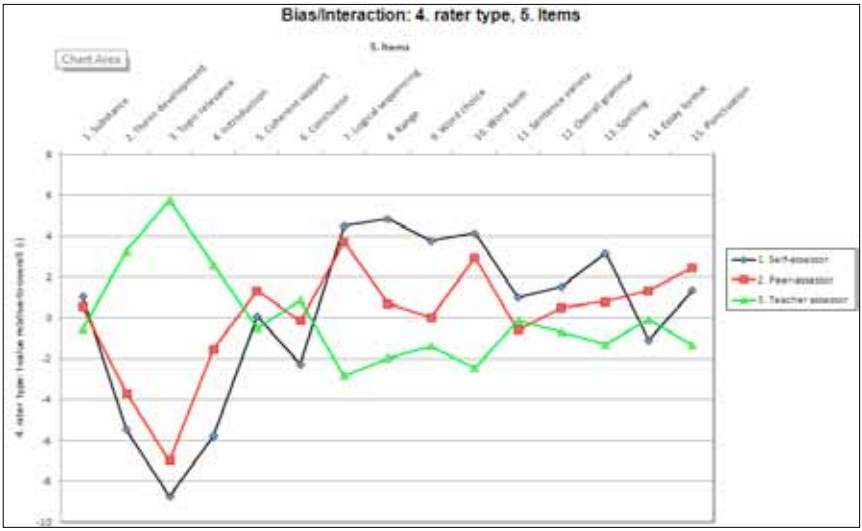


Figure 1. Bias Analysis for Rater Type (Rater-Type-Items Interactions)

Table 2. Frequency of Rater-Type-Items Bias Interactions

Item number	Items	Logit	Self	Peer	Teacher	Total	Lenient/ Severe
2	Thesis development	.09	1L	1L	1S	3	2/1
3	Topic relevance	.29	1L	1L	1S	3	2/1
4	Introduction	.07	1L	0	1S	2	1/1
6	Conclusion	.07	1L	0	0	1	1/0
7	Logical	-.36	1S	1S	1L	3	1/2
8	Range	.14	1S	0	0	1	0/1
9	Word choice	.61	1S	0	0	1	0/1
10	Word form	-.24	1S	1S	1L	3	2/1
13	Spelling	-.65	1S	0	0	1	0/1
15	Punctuation	.18	0	1S	0	1	0/1
Total	10		9	5	5	19	9/10

Note. L = Lenient, S = Severe

As shown in Table 2, which is derived from the information shown in Table 1 and Figure 1, one interesting result is that student raters (self and peer) show the opposite pattern of severity and lenience as compared with the teachers. Students are lenient for items 2, 3, and 4, whereas teachers are severe. However, the opposite is true for items 7 and 10, where students are severe, but teachers are lenient. Both students and teachers have a roughly equal division of severe and lenient interactions with the items, though all three rater types have slightly more severe than lenient bias: 5S/4L for self-assessor, 3S/2L for peer-assessor, and 3S/2L for teacher assessor.

To answer the second research question (How do self-assessors, peer-assessors, and teacher assessors differ in severity or leniency in relation to students?), another bias/interaction analysis similar to rater-items bias analysis was specified in *Facets* for rater type and students. Table 3 shows the cases of rater-type-by-students bias analysis (due to space limitation, we have included only a small part of the table). The SEs, while low, are much greater than in the rater-items bias analysis, especially for student raters. One cogent reason for the low SEs of teacher assessors is that the total number of teacher ratings across all students far exceeds that of either self-assessors or peer-assessors, because student assessors only rated one student each.

Table 3. Rater-Type-Students Bias/Interaction Analysis

Rater type	Logit	Students	Logit	Obs score	Exp score	Obs-Exp score	Bias size	Model SE	t-score	Infit MnSq	Outfit MnSq
Peer	.05	2	.47	82	63.7	1.22	-1.30	.36	-3.61	.7	.7
Peer	.05	3	.74	83	69.2	.92	-1.16	.38	-3.03	.5	.5
Self	-.17	7	.19	42	62.3	-1.35	.83	.21	4.05	.8	.8
Peer	.05	7	.19	71	57.4	.91	-.65	.24	-2.69	.5	.6
Self	-.17	8	.54	79	69.3	.65	-.69	.31	-2.22	.9	.8
Peer	.05	8	.54	82	65.2	1.12	-1.23	.36	-3.42	1.0	.9
Teacher	.12	8	.54	355	381.5	-.29	.19	.08	2.32	.7	.7
Peer	.05	11	.82	57	70.6	-.91	.64	.20	3.18	.5	.5
Peer	.05	13	.39	42	62.0	-1.33	.82	.21	3.98	.9	.7
Peer	.05	16	.21	57	45.7	.94	-.68	.27	-2.48	.9	.8
Self	-.17	17	1.43	67	75.3	-.59	.72	.26	2.80	.6	.6
Peer	.05	19	.59	79	66.3	.85	-.85	.31	-2.73	1.2	1.1
Peer	.05	24	.54	78	65.2	.85	-.81	.30	-2.71	.8	.8

Fixed (all = 0) chi-square: 1229.7 *df*: 472 significance: $p < .00$

The mean square fit statistics are good, but, unlike rater-type-items bias analysis, there are misfits in rater-type-students bias analysis. A closer inspection of Table 3 shows that out of 19 misfitting rater types, seven belong to self-assessors, 11 to peer-assessors and one to a teacher assessor. It is interesting to note that only one, the teacher assessor, is a case of underfit, and the rest are cases of overfit. Further inspection shows that student logits range from -0.07 to 0.66 for student raters, from -0.07 to 0.56 for self-assessors, and from 0.02 to 0.66 for peer-assessors, and for the teacher assessor the student is a high-ability one with a logit of 1.02. As noted in the literature, underfitting elements show noise, denoting inconsistency (Wigglesworth, 1993, 1994), and overfitting elements show lack of variation, denoting over-predictability (Linacre, 2004). Furthermore, underfitting elements are much more of a problem than overfitting ones (McNamara, 1996). When it comes to deciding how to best deal with either underfit or overfit in an existing data set, Wright, Linacre, Gustafson, and Martin-Lof (1994) claim we should first treat underfits because they force elements to remain below 1. Linacre also recommends retaining overfitting elements in the analysis because, although they do not reveal anything new, at least they tell us something. Besides, due to the lack of students' experience in rating, from a pedagogical point of view it is best to keep overfitting elements to shed light on student rating in classroom settings. Since this is not a validation study to refine an instrument, but rather a study of rater effects, by deleting misfitting elements a good deal of useful information would be lost. Considering the above-mentioned reasons, we decided to let overfitting elements stand as they are. Due to the low number of bias interactions for teacher assessors, we also did not drop the one misfitting teacher assessor.

Out of 472 bias terms, 91 were significant, with *t*-scores either greater than +2 or smaller than -2. Forty-six of the significant interactions are negative (showing leniency), and 45 are positive (showing severity).

Table 4 shows the rater-type-students bias/interaction relationship. To show the relationships, we divided students into four ability groups ranging from -0.35 to 1.45 logits. Across the top of the table is the student logit range, from the lowest ability, -0.35 logits, to the highest, 1.45 logits. Below that is the number of students in each ability group. Finally, the table shows the number of bias interactions for each rater type, divided into severe and lenient ratings.

Table 4. Frequency of Rater-Type-Students Bias Interactions

Student logits	-0.35 to -0.1	0.00 to 0.49	0.50 to 1.00	1.01 to 1.45	Total
n of students	18	105	58	7	188
Severe/Lenient	S/L	S/L	S/L	S/L	S/L
Self	1/2	9/12	7/7	1/0	18/21
Peer	2/1	16/8	5/11	0/1	23/21
Teacher	0/1	2/2	1/1	1/0	4/4
Total	3/4	27/22	13/19	2/1	45/46

Note. L = Lenient, S = Severe

As can be seen in the table, students only have a spread of 1.80 logits, and the majority (164) clusters above the mean (between 0.00 and 1.00). Only 18 students fall below the mean at the lower end of the logit scale, and only seven fall above 1.00 at the upper end of the scale. This could be attributed to the effect of the instruction, in which the students were taught the principles of essay writing, resulting in generally higher ability levels.

The majority of significant bias interactions, 81 out of 91, fall between 0.00 and 1.00 logits, while 10 occur at the extreme ends of the scale. This reflects the fact that the majority of students are clustered just above the mean, with only a relatively small number falling at the lower and upper end of the scale (seven at the lower and three at the upper). Another noteworthy point is that 45 bias interactions are severe and 46 are lenient, showing a roughly equal amount of severe and lenient bias. The third point concerning the table is that rater type shows slightly more lenient bias toward students between 0.00 and 1.00 (41 vs. 40), though again this is roughly equal. The same pattern holds true for rater type bias towards the lowest ability group (4 vs. 3). But when it comes to the highest ability group, the reverse is the case. Rater type seems to be slightly more severe rather than lenient (2 vs. 1). Of course, the low number of bias interactions at the extreme ends makes generalization difficult. There are only seven bias interactions in the lowest ability group and even fewer (only three) in the highest ability groups.

When we compare individual rater types, some interesting patterns emerge. Self-assessor and teacher assessor almost always show more or less the same pattern. For the highest and lowest ability groups, where self-assessor is lenient, teacher assessor is also lenient, and where self-assessor is severe, teacher assessor is also severe. When peer-assessor and teacher assessor are compared, the reverse is true. Where peer-assessor is severe,

teacher assessor is lenient and vice versa. Again this pattern holds true for the lowest and highest ability groups. When self-assessor is compared with peer-assessor, they mostly show the opposite pattern. When self-assessor is lenient, peer-assessor is severe, and when self-assessor is severe, peer-assessor is lenient. Although the small number of cases makes generalization difficult, it seems that self-assessor and teacher assessor ratings resemble each other more than peer-assessor and teacher assessor ratings, or self-assessor and peer-assessor ratings. This finding runs counter to Matsuno (2009) who concluded that "self-assessment was somewhat idiosyncratic and therefore of limited utility as a part of formal assessment" (p. 75).

Overall, self-assessors seem to be the most leniently biased toward students, which is in line with Ross (1998) and Matsuno (2009), who claim that students usually tend to overrate themselves. Peer-assessors are slightly more severely biased toward students, which is in line with Handrahan and Issacs (2001), who also found that peers could be very critical, but teacher assessors show severe and lenient bias in equal measure.

Discussion

The present study used MFRM to investigate bias interactions between three rater types versus first students and then items, and whether these interactions displayed systematic patterns. It further intended to argue for a place for student raters in essay rating in higher education. The findings did discover some recurring patterns. Two types of bias were found: rater type by items and rater type by students. These are explained in detail below.

Student raters (self and peer) show a pattern of severity and lenience toward items that is opposite to that of the teachers. Student raters are lenient for items 2, 3, and 4, whereas teachers are severe. However, the opposite is true for items 7 and 10, where students are severe but teachers are lenient. When we separately analyzed the data for self-assessors, peer-assessors, and teacher assessors, teacher assessors were different from student assessors. The most likely explanation for this is that student assessors were monitored while they were rating the essays while teacher assessors were not. They were rating on their own and they might have had their own interpretations of the criteria, as is quite common in the literature (Lumley, 2005). The monitoring influenced student assessors to have similar rating patterns to each other. This has been shown to result in consistency (see, for instance, Knoch, Read, & von Randow, 2007).

Both self-assessors and teacher assessors show the opposite pattern of severity and leniency as compared with peer-assessors toward the extreme ends of student ability groups. Unlike Kondo-Brown (2002) and Schaefer (2008), who found that raters tended to have more severe or lenient bias toward the extreme ends of ability groups, the present study found that the rater type tended to have more lenient or severe bias patterns toward the midpoints of ability groups, which, as was mentioned in the results section, could be attributed to the instruction students received, making them cluster around the mean, thereby attracting rater type. Like Kondo-Brown's and Schaefer's studies, the present study also found that rater type could show more severe bias toward the highest ability students and more lenient bias toward the lowest ability students. This might be because of the rater type's raised expectations toward the highest ability students or because rater types gave the benefit of the doubt to the lowest ability students, as Schaefer argues, or it might be simply because of the *Facets* program's inability to accurately estimate ability levels at the extreme ends of the continuum, as Kondo-Brown maintains.

Self-assessors tend to have the most severe bias toward items and peer-assessors seem to have the most severe bias toward students. Severity of peer-assessors toward students is mainly because they tend to be critical of their peers and is in line with many previous studies including Handrahan and Issacs (2001). The findings of the present study also corroborate Matsuno (2009) in that peer-assessors in the present study also showed fewer bias patterns toward items, compared to self-assessors. Self-assessors' larger number of bias patterns toward items may be because they did not have a clear understanding of the assessment criteria.

Spelling is the easiest item as scored by rater type. This finding is consonant with Matsuno (2009) and Kondo-Brown (2002) and it is because superficial features like spelling are usually not given in-depth thought (Hamp-Lyons, 2003). It is also in line with Mahoney (2011) who, in the context of error gravity in the Japanese context, asserted that spelling is not as important as other language elements. Word choice is the most difficult item as scored by rater type. This finding runs counter to many previous studies including McNamara (1996), Lumley (2005), Matsuno (2009), and Schaefer (2008). A possible reason could lie in relation to the setting in which the respective studies were done. It seems that different studies in different settings using different raters produce different results concerning the most difficult items and this could be attributed to the perceptions, experiences, and cultural inclinations of raters. McNamara's study was done in an ESL

setting, using highly trained professional raters, and those of Schaefer and Matsuno were done in EFL settings, the former using rather inexperienced native English-speaking raters and the latter using student raters. Another possible interpretation, as it relates to the present discussion and as has been confirmed in previous studies (Saito & Fujita, 2009), might be that raters, especially student raters, are generous in their rating of some items or are unable to differentiate items, hence resulting in inflated marking.

The present study is inconclusive as to whether self- or peer-assessment could be an alternative to teacher assessment in awarding grades on essay writing. There are some inconsistencies. As seen in Table 2, both self-assessors and peer-assessors rate very similarly to each other. In most cases, where self-assessors are lenient, peer-assessors are also lenient and where self-assessors are severe, peer-assessors are also severe. These patterns run counter to teacher assessors who have the opposite pattern. Table 4, however, reveals a different pattern. Here self-assessors and peer-assessors rate mostly differently, and self-assessors rate similarly to teacher assessors. Self-assessors tend to overrate themselves, a finding that is consistent with previous research in which low ability students tend to overrate themselves and high-ability students tend to underrate themselves (Blanche & Merino, 1989; Boud & Falchikov, 1989), which could be attributed to experience (Ross, 1998), subjective points of view such as habits of overestimating of self-ability (Saito & Fujita, 2004), or cultural values of modesty and ego (Brown, 2005). In Iran, evaluation is norm-referenced. When assigning grades, teachers routinely compare a student's work to other students' work (Farhady & Hedayati, 2009). Consequently, when Iranian students rate their own essays, they do not tend to assign ratings that are lower than those that they would assign to their peers' essays. Because students very much appreciate higher ratings, they may be more likely to assign their own essays higher ratings than they actually deserve. Student raters are, however, consistent when it comes to assessment criteria, which is in keeping with Falchikov and Goldfinch (2000), who conclude that when the criteria are explicitly stated and well understood, they lead to more accurate and consistent marking by student raters. The inconsistencies in the present study are partly because the nature of self-assessment and peer-assessment is not yet known and more research is needed to show their efficacy in L2 testing. For example, Saito and Fujita (2004) argue that "lack of research on the characteristics of peer-assessment in EFL writing may inhibit teachers from appreciating the utility of this innovative assessment" (p. 31). Another plausible interpretation for the inconsistency of the results could be the lack

of research using MFRM in this area. As Matsuno (2009) states, "as more researchers use this research technique, we can illuminate a multitude of facets of self and peer assessments" (p. 95). Lack of any meta-analytic study in which findings of other studies are aggregated to arrive at a consensus may well be another reason for the inconsistency.

This study has possible implications for rater training. One hour of training coupled with monitoring in the present study led to more consistency on the part of student raters. Although rater training may not eliminate rater error, it could lead to consistency, especially when it is combined with monitoring. In cases such as this study in which students are involved in rating essays and are going to share rating with teachers in language settings, it is best to provide them with enough training and monitoring. Although the findings in the present study are inconsistent as to the similarity between self, peer, and teacher rating, it was shown that self and teacher ratings were similar to each other, which provides partial evidence for the concurrent validity of self and teacher ratings.

This study is limited in many ways. First, it was purely quantitative. Adding a qualitative component could provide deeper insights into why such findings were obtained. The second limitation relates to the small number of teacher assessors in this study, although the number of teacher assessors was greater than in other studies in which self- and peer-assessments were involved. Most published studies employing self-assessment, peer-assessment, and teacher assessment have used only a very small number of teacher assessors. It would be good for future studies to use a larger number of teacher assessors along with self-assessors and teacher assessors in different settings to see if the results would be the same or different. The third limitation is the small number of essays both self-assessors and peer-assessors rated (one essay each). This is also a further avenue for research, in which future studies could have both self-assessors and peer-assessors rate a larger number of essays because, as was shown in this study, a larger number of ratings could lead to smaller error, which might reduce the number of biases. Finally, due to the small sample size we cannot really make any statements about whether self-assessment or peer-assessment could be a reliable alternative to teacher assessment. The rater bias subpatterns are also based on small differences and need to be interpreted with caution. Researchers should strive to answer this important question in future studies.

Conclusion

Differential rater severity or bias effect as a pervasive rater effect is detrimental if not detected and treated appropriately. In the present study, all three types of rater had bias patterns toward either students or items; furthermore, although these bias patterns were more or less similar, it seems that they were also unique in that each rater type had a particular bias pattern.

Such differences, especially those between students and teachers, seem to be inevitable because they are also manifested in other EFL or ESL settings and confirm previous studies which empirically showed that rater training may reduce rater errors or may make raters self-consistent, but does not necessarily eliminate rater errors (see Knoch, 2011). There are a few methods to help reduce bias when detected. One is to provide rigorous training coupled with monitoring, but the optimal type and amount of training is yet to be shown. In our study, one hour of training led to the consistency of student raters; still, there were many cases of bias, which suggests that it was not enough. Another helpful way is instruction, which might dispense with the need for rater training (Saito, 2008). Instruction might provide raters with clear and explicit assessment criteria or might involve co-creation and negotiation of rating scales with raters. In our study, raters were not provided with such instruction, which might be another reason for bias. Feedback to raters has also proved to be helpful, especially if it is longitudinal (Knoch, 2011). Another reason for the presence of bias in the present study might be lack of feedback.

Coniam (2008) noted that the use of scoring rubrics in performance testing is now common across Asia. The use of peer- and self-raters is also something that is attracting attention in Asia and beyond. MFRM has proven to be useful in researching and testing in this area, and we hope that this study has contributed to the development of this area in EFL settings.

Acknowledgement

We would like to acknowledge the helpful advice of Mike Linacre and Carol Myford regarding the use of the *Facets* computer program to analyze the data.

Farahman Farrokhi is an Associate Professor teaching and supervising MA and PhD students at the University of Tabriz. His areas of interest include classroom management and assessment.

Rajab Esfandiari is an Assistant Professor at Imam Khomeini International University in Qazvin, Iran. His main areas of interest include Rasch measurement.

Edward Schaefer is a Professor in the Graduate School of Humanities and Science at Ochanomizu University. His areas of interest include Rasch measurement and L2 writing instruction and assessment.

References

Bailey, E. P., & Powell, P. A. (2008). *The practical writer with readings*. Blake, VA: Thomson Wadsworth.

Blanche, P., & Merino, B. J. (1989). Self-assessment of foreign language skills: Implications for teachers and researchers. *Language Learning, 39*, 313-340.

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education, 18*, 529-549.

Brown, A. (2005). Self-assessment of writing in independent language learning programs: The value of annotated samples. *Assessing Writing, 10*, 174-191.

Coniam, D. (2008). An investigation into the effect of raw scores in determining grades in a public examination of writing. *JALT Journal, 30*, 69-84.

Du, Y., Wright, B. D., & Brown, W. L. (1996, April). *Differential facet functioning detection in direct writing assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*, 155-185.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H).* Strasbourg, France: Council of Europe/Language Policy Division. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/CEF-refSupp-SectionH.pdf

Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*, 171-191.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93-112.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment: A meta-analysis comparing peer and teacher remarks. *Review of Educational Research, 70*, 287-322.

Farhady, H., & Hedayati, H. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics, 29*, 132-141.

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4*, 113-148.

Fukazawa, M. (2010). Validity of peer assessment of speech performance. *Annual Review of English Language Education in Japan, 21*, 181-190.

Hamp-Lyons, L. (1991). Scoring procedures in ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Norwood, NJ: Ablex.

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 162-189). Cambridge: Cambridge University Press.

Handrahan, S., & Issacs, G. (2001). Assessing self- and peer-assessment: The students' views. *Higher Education Research and Development, 20*, 53-70.

Holster, T. A. (in press). Many-faceted Rasch analysis of student peer assessment. *Studies in the Humanities*.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior—a longitudinal study. *Language Testing, 28*, 179-200.

Knoch, U., Read, J., & von Randow, T. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing 12*, 26-43.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese L2 writing performance. *Language Testing, 19*, 3-31.

Linacre, J. M. (1989/1994). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (2004). Optimizing rating scale effectiveness. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258-278). Maple Grove, MN: JAM Press.

Linacre, J. M. (2011). *FACETS* (Version 3.68.1) [Computer Software]. Chicago, IL: MESA Press.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt am Main: Peter Lang.

Mahoney, S. (2011). Exploring gaps in teacher and student EFL error evaluation. *JALT Journal, 33*, 107-130.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*, 75-100.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.

Meyers, A. (2006). *Composing with confidence: Writing effective paragraphs and essays.* New York: Pearson Longman.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In E. V. Smith, Jr. & R. M Smith (Eds.), *Introduction to Rasch Measurement* (pp. 460-517). Maple Grove, MN: JAM Press.

Negishi, J. (2010). Multi-faceted Rasch analysis for the assessment of group oral interaction using CEFR criteria. *Annual Review of English Language Education in Japan, 21*, 111-120.

O'Neill, T. R., & Lunz, M. E. (1996, April). *Examining the invariance of rater and project calibrations using a multi-facet Rasch model*. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (rev. ed.). Copenhagen: Danish Institute for Educational Research.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*, 1-20.

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing, 25*, 553-581.

Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research, 8*, 31-54.

Saito, H., & Fujita, T. (2009). Peer-assessing peers' contribution to EFL group presentations. *RELC Journal, 40*, 149-171.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*, 465-493.

Smalley, R. L., Ruetten, M. K., & Kozyreve, J. R. (2000). *Refining composition skills: Rhetoric and grammar*. Boston: Heinle & Heinle.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*, 305-335.

Wigglesworth, G. (1994). Patterns of rater behaviour in the assessment of an oral interaction test. *Australian Review of Applied Linguistics, 17*(2), 77-103.

Wright, B. D., Linacre, J. M., Gustafson, J.-E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370. Retrieved from http://rasch.org/rmt/rmt83b.htm

Yamanishi, H. (2004). How are high school students' free compositions evaluated by teachers and teacher candidates? A comparative analysis between analytic and holistic rating scales (in Japanese). *JALT Journal, 26*, 189-207.

Appendix

Essay Rating Sheet

Essay number:

Rater's name:

	Very poor	Poor	Fair	Good	Very good	Excellent
1. Substance	1	2	3	4	5	6
2. Thesis development	1	2	3	4	5	6
3. Topic relevance	1	2	3	4	5	6
4. Introduction	1	2	3	4	5	6
5. Coherent support	1	2	3	4	5	6
6. Conclusion	1	2	3	4	5	6
7. Logical sequencing	1	2	3	4	5	6
8. Range	1	2	3	4	5	6
9. Word choice	1	2	3	4	5	6
10. Word form	1	2	3	4	5	6
11. Sentence variety	1	2	3	4	5	6
12. Overall grammar	1	2	3	4	5	6
13. Spelling	1	2	3	4	5	6
14. Essay format	1	2	3	4	5	6
15. Punctuation/capitalization/handwriting	1	2	3	4	5	6

Reviews

Global Englishes in Asian Contexts: Current and Future Debates. Kumiko Murata and Jennifer Jenkins (Eds.). Hampshire, UK: Palgrave Macmillan, 2009. ix + 233 pp.

Reviewed by
James Essex
Komazawa University

English, despite being an oft-used word, is still listed by a number of dictionaries as uncountable, and indeed, if one tries to pluralize it in MS Word, it is highlighted as an error. With the number of people who speak English as a foreign language (EFL) and English as a second language (ESL) now outnumbering those for whom it is a first language (Crystal, 1997), this book offers valuable insights into the spread of English into various Asian contexts and the phenomenon known as World Englishes (WE). *Global Englishes in Asian Contexts: Current and Future Debates* provides a concise discussion of the use of English in Asia (an ever expanding usage region which sometimes includes Middle Eastern countries), New Zealand, and Australia. Divided into four parts, comprising 12 essays written by well-known researchers in the field, this comprehensive book is an ideal springboard for research and discussion.

Part I, "Understanding Englishes and English as a Lingua Franca in Asia" defines the notions of WE, English as an International Language (EIL), English as a Lingua Franca (ELF), and their inter-relationships. The idiom principle—the combination of words in phrases in the interests of effective communication—is covered extensively, and turns out to be the unifying theme of Part I. Such phrases not only help to facilitate shared meaning but also serve to establish rapport and to identify speakers as members of the in-group, and are thus markers of shared territory. The idiom principle attempts to account for language which does not fit what Sinclair (1991) calls the "open choice principle." The idiom principle illustrates that semi-preconstructed phrases exist, which constitute single choices, even though

they seem to be analyzable into segments. In her essay related to the idiom principle, Jenkins describes the notions of accommodation and code-switching, which she argues are often performed by ELF speakers in order not only to compensate for gaps in knowledge but also to signify group membership and solidarity. She concludes by claiming that in the future, people occupying the top of the English language hierarchy will not be native speakers (NS) but bilingual speakers "who have the skills to function comfortably in multilingual situations" (p. 52). Linked succinctly to this idea is Smith's discussion of a recurring theme in WE: understanding across cultures. He discusses the notions of comprehensibility, intelligibility, and interpretability and argues that all three are crucial if cross-cultural understanding is to occur. The chapter ends with several practical suggestions on how to improve understanding in WE contexts. The first part of the book raises some very interesting questions, namely that of whether or not nonstandard varieties should be taught, answers to which can be found in Part II.

In Part II, "Cultural Identity, Ideology, and Attitudes in English in Asia," Hung's exploration of how Shaw's *Pygmalion* might have been written in Singaporean English highlights the standard-vs.-nonstandard debate and turns out to be the *pièce de résistance* of the book as it encompasses the questions the book seeks to answer: Should we, as language teachers, teach a standard variety of English or should we expose learners to nonstandard varieties? Will regional varieties of English suffice? The answer unfolds in the next eight chapters, and the unequivocal answer is that we must pay heed to both the culture and the learning context of our learners. At first sight, Hung's essay is a subtle attack on the Singaporean government, which opposes the use of Singlish, a nonstandard variety of Standard Singapore English. It serves, however, as an attack on any government or institution which seeks to prescribe a certain language variety. Mamoru places this in the Japanese context in Chapter 6 with an examination of Japanese English for EIAL (English as an International Auxiliary Language), wasting no time in explaining the necessity that nonnative speaker (NNS) English be clearly differentiated from NS English. Similarly, Park argues her case from a Korean culture-specific perspective, predicting that not too far in the future Korean English will become a recognized glocalized (globalization + localization) variety. As such, she advocates the need to teach this variety not only to Korean English learners but to business people from outside Korea. She concludes by saying that certain forms, phrases, grammar, and sentences should be regarded as norms as long as they are understandable (p. 97). Essentially, importance should be placed on making oneself understood.

Part III, "Englishes in Asian Academic and Business Contexts," could be of value to teachers involved in teaching scientific writing, business English, and English in advertising. Indeed, the unifying thread of this unit is English in academic and business contexts. Yamuna Kachru looks at academic writing in WE in the Asian context, arguing for the need to recognize the cultural conventions and values found in scientific writing, which she claims are not necessarily universal. Gill discusses standards and realities of English in the Malaysian workplace, and Bhatia discusses English in Asian advertising and in the teaching of WE, highlighting the use of English found in advertising.

The fourth and concluding part, "The Future of Englishes: A Paradigm Shift?" discusses Asian Englishes in the Asian age and the challenges faced by users of English in the Asian region, principally India and China. Pennycook discusses what he calls plurilithic Englishes, or the avoidance of regional reproduction of authority, and power of NS and, conversely, the lack of integrity and comprehensibility. After examining Kachru's well-known model of concentric inner, outer, and expanding circles making up WE, Pennycook concludes by proposing a more fluid model which includes language users and their individual contexts. Finally, Yano discusses the future of English and the need to move beyond the Kachruvian three-circle model, describing a new wave of regional standard Englishes (RSEes). He illustrates how English in Asia has evolved and will continue to do so.

One of the book's strengths is its accessibility; the arguments, while seemingly complex and multifaceted, are easy to understand. However, some chapters add nothing new to the existing literature. In a field which is already saturated, one cannot help but wonder whether this book will be lost among those which offer a less diluted discussion of specific contexts. Personally, I found Mamoru's chapter, "Japanese English for EIAL," to be informative and useful because it is all about context, echoing the notion that not only should we pay attention to learners' culture and learning context, but also that Asia may be too large to cover comprehensively in only one volume. However, this book does exactly what it sets out to do in the Introduction from Murata and Jenkins. It provides readers with a good understanding of the spread of English throughout Asia.

References

Crystal, D. (1987). *The Cambridge encyclopedia of language*. Cambridge: Cambridge University Press.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Young Learner English Language Policy and Implementation: International Perspectives. Edited by Janet Enever, Jayne Moon, and Uma Raman. Reading, UK: Garnet, 2009. 248 pp.

Reviewed by
William Green
Sapporo University

Over the past 30 years, teaching English to young learners (TEYL) in primary schools has become increasingly common all over the world. Governments recognize that globalization means the labour force will need foreign language skills in order to compete in the international marketplace. The introduction of compulsory foreign language classes for Japanese fifth and sixth grade pupils in April 2011 is a further example of this change.

Young Learner English Language Policy and Implementation: International Perspectives, edited by Enever, Moon, and Raman, is the proceedings of a conference held in Bangalore in 2008 which aimed to facilitate the sharing of experiences, broaden perspectives, and influence future policies in the introduction of English as a foreign language in primary schools. The book is organized into three sections: The first section includes the three keynote presentations from the conference; the second presents 12 national case studies of implementations of early English programmes; and the third section includes 12 smaller scale innovations, experiments, and projects. In addition, the editors have written an excellent introductory chapter that provides an overview of the current challenges and issues in TEYL. The chapter ends with a list of 12 recommendations for developing policy and implementing English as a foreign/second language in primary schools.

Johnstone's keynote (Chapter 3) expands on the introductory chapter by providing a useful set of conditions for success in TEYL programmes. Johnstone makes an interesting but somewhat controversial point about including reading, writing, and grammar at age 7 or 8. This, he argues, helps children think analytically and strategically in their learning. An associated idea is to raise learners' awareness of their success, which increases their motivation and encourages them to monitor their own learning. Together, the introduction and this chapter provide an excellent overview of and potential solutions to problems facing policy-makers and practitioners in TEYL.

The chapters in the rest of the book describe challenges that have arisen once innovation has been implemented and the ways in which these challenges have been addressed. The book poses an implicit question about what the role of each stakeholder—local education authorities, principals and teachers, teacher-trainers, external agencies, and organizations—should be, and how they should collaborate to create successful English teaching in primary schools.

As Goto Butler notes in Chapter 2 about Korea, Taiwan, and Japan, a typical challenge is the mismatch between a curriculum influenced by Western thinking (e.g., communicative or task-based language teaching) and the local educational culture. For example, Turkey's revised curriculum introduced English from Grade 4 in 2006 and is described as being constructivist and learner-centred (Chapter 8). However, a survey conducted with 52 primary English teachers in 2007 showed that classes were still largely teacher-centred and a constructivist approach had not been adopted. This mismatch is echoed in other countries such as India (Chapter 9) and China (Chapter 15).

Teachers clearly need help to adjust their classroom approach if major changes to the English curriculum are not based in local educational culture. Many of the chapters end with a call for more and improved teacher training. In contexts where textbooks are produced or selected by central government, teachers need long-term training in methods suitable for implementing the books in ways which are consonant with their existing practices; their training is more likely to be successful if it includes such an element of cultural continuity.

A further key responsibility for teacher trainers is to provide student-teachers with problem-solving approaches such as reflective practice, action research, and exploratory practice. These approaches give teachers a way of examining issues in their own teaching and arriving at solutions that may help them bridge the gap between the imposed curriculum and their classroom reality. As part of pre-service training, this would fall under the remit of university lecturers who, in their role as researchers, could also usefully conduct studies in local schools. For example, a small pilot study in Oman, described in Chapter 17, assessed a primary English reading programme and improved it by adding a phonics element. In Turkey, a case study approach to innovations in the primary English curriculum used a questionnaire, classroom observations, and interviews to provide a profile of teachers, on which to ground suggestions for improvements to the system (Chapter 22).

The idea that materials can help to train teachers in TEYL methods emerges in a number of papers. The teacher's book accompanying the students'

course book was redesigned in Oman to include tasks that teachers could combine at their own discretion (Chapter 17), thus facilitating more teacher decision-making and control over lesson planning. In the publishers' forum, reported in Chapter 25, it was argued that materials may help train teachers by providing rationales for activities and not just the activities themselves. Such materials design allows minimally qualified teachers to teach English while encouraging their decision-making, autonomy, and development, and means that teachers with basic English ability can maximize their students' output.

Teachers can also be supported in a variety of social interactions; a number of interesting examples are described. In Cameroon, for example, the national English teachers' association compensated for a poorly planned and under-funded national bilingualism policy by running workshops in primary schools to help Francophone teachers teach English (Chapter 10). In Nigeria, a series of radio programmes provided teacher development to practitioners who would otherwise have received none; the programmes had an estimated 70 million listeners (Chapter 23). This was one of the very few chapters to deal with technology in TEYL, and this is an area which could easily be further developed by teachers' associations, universities, and others, particularly as a way of supporting teachers.

There are a number of recurring challenges in introducing TEYL programmes, for example, teachers' abilities in English and familiarity with suitable teaching techniques, training teachers and teacher-trainers, curriculum design, materials, and assessment. This book illuminates many of these areas with fascinating accounts of innovations from around the world. Teachers of fifth and sixth grade pupils across Japan are now teaching English to their students, perhaps for the first time. Although this book provides little for them in the way of direct assistance, it is hoped that teacher educators will draw inspiration and ideas from the varied English language education policies and programmes presented.

EAP Essentials: A Teacher's Guide to Principles and Practice. Olwyn Alexander, Sue Argent, and Jenifer Spencer. Reading, UK: Garnet, 2008. viii +379 pp.

Reviewed by
Rob Higgins
Kwansei Gakuin University

EAP Essentials is written by EAP (English for Academic Purposes) practitioners for EAP practitioners. It has not been written as a resource tool for academic research, though there are enough references to satisfy this demand. It is also not a pedagogical tool to take to the classroom for clarification during teaching time. In the introductory material, the aims and intentions of this text are outlined. These aims include bridging the gap between theory and practice, emphasizing reflective practice, and providing the same kind of support and guidance that you might gain from working alongside an experienced EAP teacher.

The book has 10 chapters: "The Context of EAP," "Text Analysis," "Course Design," "Reading," "Vocabulary," "Writing," "Listening and Speaking," "Critical Thinking," "Student Autonomy," and "Assessment." The concept of reflective practice, which suggests that learning through action is a continuous process, acts as a framework for interacting with this text. Each chapter provides hands-on tasks to be completed by the reader. This encourages reflection on what has been introduced and, moreover, allows the reader to contextualise this approach with personal experiences. The end of each chapter gives several relevant references that can be pursued later. There is also a CD-ROM containing photocopiable resources for classroom activities and teachers' notes, which are conveniently signposted throughout the text where the activities relate to the discussion and theme.

The practical tone of the book, grounded in EAP discourse, focuses on the relationship between EAP teachers and students. Actual comments and reflections of anonymous students are included, reinforcing the reflective dimension of each chapter, and suggesting that EAP practitioners carefully consider the specific academic requirements of their students. EAP has a long tradition of focusing on writing to raise awareness of academic conventions and enable students to prepare for specific discourse communities (Tribble, 2009). The field has not, however, clarified as successfully other methodological aspects of EAP. For example, it has been suggested that EAP

has focused on *what* needs to be taught and *why* it needs to be taught with less emphasis on *how* it needs to be taught. The authors bring this discussion to a practical level with concrete examples of appropriate materials and the *how to* of designing materials (Watson-Todd, 2003).

Chapter 3 develops the salient theme of focusing on the practical teaching context in EAP through needs analysis while providing a context specific framework. By identifying the different constraints on course design and needs, the authors suggest that decisions regarding course design should rest firmly in the hands of the teachers. The authors broaden this discussion by exploring the *how* of EAP-context specific methodology, with the purpose of designing a coherent syllabus. Case studies and authentic teaching materials help readers evaluate and reflect on this process throughout the chapter.

The next part of the book examines the four skills and vocabulary. Each chapter balances the academic research dimension with suggested activities and case studies related to the skill. In the reading section, for example, the complexities of research article abstracts are considered. The book does not shy away from difficult and complex areas of EAP and examines these topics through a process of case study analysis and follow-up reader activities. In the vocabulary chapter, the authors highlight the necessity of teaching both general academic vocabulary and subject-specific vocabulary.

Writing (Chapter 6) is, as the authors acknowledge, "the most crucial of the skills needed in an academic context, where written texts are the main means of communication" (p. 178). This section highlights the expectations of content teachers (who are not ESL/EFL teachers) in terms of what they want students to be able to perform in relevant disciplines. The analysis is developed through actual comments about the difficulties students have when doing writing for the first time, which include "I don't know the correct academic style" and "A dissertation for me is a completely new thing" (p. 178). Students' comments about teachers' expectations of student writing at the university level, together with actual comments from teachers, add relevance to the discussion and support its focus.

The listening and speaking chapter provides a detailed analysis of academic lectures. It discusses the complexities of lectures; this approach is helpful in building an understanding of the structure of lectures in particular disciplines. These complexities are addressed in the designing of listening tasks, which promote awareness raising and developing bottom-up and top-down listening strategies to support and overcome linguistic difficulties during lectures. The chapter also focuses on developing skills to recognise content relationships in lectures.

The final three chapters of the book look at critical thinking, student autonomy, and assessment. These are a welcome addition to the discussion of EAP course design and relate to the overall focus on academic study skills developed in the book. Throughout these chapters, the authors acknowledge the importance of reflective teaching and learning as a means to understand and respond to the challenges of academic contexts. The emphasis on reflection supports a general recognition in ELT that practitioners must analyse their own teaching and students must analyse their own learning.

EAP Essentials has much to offer one who has little formal training in EAP teaching. It explains relevant theoretical aspects of EAP for less experienced teachers and delivers on its aims to link theory and practice through reflection. Furthermore, through the support and guidance of experienced practitioners, the book outlines some very practical solutions to the challenges of teaching EAP. This practitioner emphasis will satisfy experienced teachers' desires to gain insights into other teaching contexts. Overall, the book is a welcome addition to a rapidly growing discipline that requires further clarification regarding approaches and methods.

References

Tribble, C. (2009). Writing academic English: A survey review of current published materials. *ELT Journal, 63,* 400-417.

Watson-Todd, R. (2003). EAP or TEAP? *Journal of English for Academic Purposes, 2*, 147-156.

The Age Factor and Early Language Learning. Marianne Nikolov (Ed.). Berlin: De Gruyter Mouton, 2009. 424 pp.

Reviewed by
Jung In Kim
SUNY at Buffalo

The importance of children's English education has been recognized widely in countries where English is learned as a second or foreign language. This compilation of research into how the age factor interacts with other factors in a variety of educational contexts offers a clear, broad perspective of the age factor in early language learning. The book comprises 17 chapters that

examine early language learning and teaching in Europe, Asia, and North America. Nikolov has put together an excellently edited volume to explain how context contributes to early language teaching and learning.

The first two chapters examine mainstream SLA research trends, giving novice teachers and researchers of young language learners a general scaffolding within which to place SLA. The next four chapters focus on the assessment of young learners. Curtain (Chapter 3) introduces a variety of approaches to assessment of early second language learning in the USA such as the Early Language Listening and Oral Proficiency Assessment (ELLOPA), the Student Oral Proficiency Assessment (SOPA), and the CAL (Center for Applied Linguistics) Oral Proficiency Exam (COPE). Comparing these common assessment instruments according to grade level, program, and test format gives readers a useful outline of different types of assessment for young language learners. In Chapter 4, Inbar-Lourie and Shohamy compare the effectiveness of two models of teaching in distinct educational contexts. The first model comprised expert EFL teachers who taught only English to young learners while in the second model, teachers were generalists or homeroom teachers. Inbar-Lourie and Shohamy suggest that teachers need to take into account a different formative assessment construct which includes not only language but also a content focus. In Chapter 5, Jalkanen investigates an early total immersion program with KATE (the Kuopio Assessment Tool for English) and concludes that early total immersion is an appropriate methodology for young language learners. These studies identified the need to use appropriate assessments depending on the levels of learners, purpose of the program, and teachers' qualifications.

Chapters 7 and 8 examine how age impacts on language learning over time. In Chapter 7, Muñoz draws on results from the BAF (The Barcelona Age Factor) project, a longitudinal study from 1995 to 2004 with the aim of exploring age-related differences and attainment of language. Her study confirmed that older language learners outperformed younger learners in the beginning stage. Furthermore, young learners did not show a long-term advantage within 9 years (1995 to 2004), which is the opposite of the prevalent findings that younger starters attain higher levels of language proficiency. Kasai (Chapter 8) investigates how age plays its role in acquiring English /l/ and /r/ sounds with Japanese children and adult learners. Results showed that whereas there is an age effect in production of those sounds, there is no age effect in discriminating between them. The critical period of language learning has been debated as a limitation of SLA. These two studies displayed that "the younger, the better" in second or foreign language learning does not seem to fit all instructional contexts.

The next two chapters look at individual differences (attitude, motivation, context, and aptitude) in early language learning. Djigunović (Chapter 9) explores attitudes, motivations, strategies, and anxieties of learners in early language programs. Strikingly, within individual differences in early language learning, the role of teachers stands out among those variables but is not considered as important as individual factors. Mattheoudakis and Alexiou in Chapter 10 highlight socioeconomic factors in early language learning. The findings show that there are significant differences between eastern and western schools. However, it is notable that several of the teachers' opinions and beliefs about children from low-income families were not confirmed by the findings. Therefore, the authors stress that teacher beliefs or expectations of certain children may contribute to students' ultimate achievement of language learning.

From Chapter 12 to Chapter 17, this book delves into the macro-level of language teaching and learning: curriculum, language policy, and language choice. Wang (Chapter 12) and Moon (Chapter 13) investigate teacher attitudes toward curriculum change in China and Vietnam respectively. Even though teachers themselves may support an innovative curriculum, prevailing conditions such as a heavy workload, large classes, and a lack of teacher training make some changes an unrealistic goal for English language learning and teaching. Peng and Zheng (Chapter 14) also point out a mismatch in language policy and the language classroom in China. Although the Chinese Ministry of Education emphasizes learners' communicative competence and the application of communication strategies, students and teachers used English very infrequently when faced with language difficulties in the classroom. Rather, students simply used a Chinese word to overcome any communication challenge. These three chapters apparently reflect the reality of second language teaching and learning in EFL settings. What teachers perceive to be an innovative teaching methodology may be limited by the broader social context, such as national curriculum and high stakes tests, rather than a teacher's self-will.

Looking across all 17 chapters, two weaknesses appear. The depth of data analysis to reach the conclusions that the authors make remains in question. Comparing research procedures and research design, many of the studies spend relatively little time discussing either the data analysis of their work or the findings. Furthermore, it would have been more practical if the authors included pedagogical implications that related to their research findings. Current and future teachers expect to read how those research findings can be used to improve their teaching practice. Therefore, researchers need to make an effort to link theory and practice.

Despite these two criticisms, this book makes a contribution not only to SLA, but also to early language learning in different contexts. Indeed, the book deals with crucial issues in SLA as well as how those issues can be interpreted for young language learners in different educational contexts. As mentioned, it is hard to find appropriate research-based books that focus on young language learners. In particular, a compilation of research in EFL contexts is not easy to find. Therefore, I would recommend this thought-provoking book to all those interested in teaching English to young learners in either ESL or EFL contexts. It is a valuable read to see what is happening in countries where foreign languages are taught to young learners.

***Sociocultural Theory in Second Language Education: An Introduction Through Narratives*. Merrill Swain, Penny Kinnear, and Linda Steinman. Bristol, UK: Multilingual Matters, 2011. xvii + 174 pp.**

Reviewed by
Tim Murphey
Kanda University of International Studies

Many of us have long been searching for an effective introduction to sociocultural theory (SCT) for our graduate and undergraduate students, to scaffold them into working conceptualizations that they can grow with. Apparently, the authors of this volume also could not find one, so they wrote it. Swain, Kinnear, and Steinman have created a user-friendly text that not only scaffolds readers gently into the essential concepts of SCT, but does so in the entertaining and captivating mode of narratives.

The book starts off with a brief narrative about Vygotsky's work and life and then an introduction dealing with SCT in SLA through narratives. Swain, Kinnear, and Steinman rightly deal immediately with the readers' question, "Why stories?" They answer thoroughly, devoting four pages to their arguments, with perhaps the most important being that we remember stories better than anything else: "Stories have the quality of 'stickiness' that lasts after discrete bits of information are forgotten" (Heath & Heath, 2007, p. xi). The narrative turn, just like the social turn in SLA, is still occurring on many campuses and across many disciplines. The authors then narrate the

process of producing the book: "We elicited/solicited the narratives in circumstances separate from the conception of this textbook, and then studied these narratives to demonstrate particular SCT concepts-in-context" (p. xiii). They also emphasize that "this is not research *on* narrative; rather it is achieving/seeking understanding of phenomena (SCT concepts) *via* narratives" (p. xiii, italics in the original).

The first seven chapters deal with some of the primary concepts in SCT, clearly marked and allowing those with particular questions about these concepts to go straight to the topic of their concern: mediation, zone of proximal development (ZPD), languaging through private and collaborative speech, everyday concepts and scientific concepts, interrelatedness of cognition and emotion, activity theory, and assessment. Each chapter starts with key terms (also glossed at the end of the book), key tenets of the primary concept as it informs SCT, and the context of the story. The main part of the chapter is of course the narrative (by a language learner, teacher, or in an interview format), followed by the authors' interpretations (something like a movie director's scene commentary on a DVD) which, as they note, are often a story about the story.

The authors end each of these chapters noting allied concepts, controversial issues, key studies, and questions and implications for research and pedagogy. All references are at the end of the book following a useful glossary. In addition to the primary concepts they also use, exemplify, and gloss many others such as scaffolding, communities of practice, affordances, agency, genesis, and the genetic method. The layout of the book is also very user friendly and attractive, especially the first page of each chapter, which gives an overview of the chapter with bullet points.

My favorite two stories were "Madame Tremblay" (Chapter 2, illustrating the ZPD) and "Grace" (Chapter 5, illustrating the interrelatedness of cognition and emotion). While the post-study analyses are extremely stimulating, the narratives as the primary material are captivating. I took to "Madame Tremblay" because it was written by a university student about her French immersion schooling in fourth grade in a vivid literary style with great humor and confusion (which may be difficult for nonnatives). "Grace," on the other hand, was a part of a 2-hour interview with a woman who had grown up as a Greek in Canada (speaking Greek at home), left at 24 to teach English in Greece for 20 years, and returned to Canada to start graduate school in her mid 40s. The sense that she was an outlier in both places and not a native speaker of either language created emotions and identity issues that many long-time-abroad speakers feel. With these stories as data,

readers also get the idea that SCT can deal with real people, problems, and possibilities-in-context.

Chapter 8, the last chapter, invites readers to participate in interpreting two narratives the way the authors did in the previous chapters. After having read the authors' interpretations in the previous seven chapters, the SCT concepts actually start jumping off the page at you; you really want to talk to someone about them and share your conceptualizations and analyses. The book ends with a short Discussion section where the authors themselves reflect on what they learned in writing the book (walking their SCT reflective talk), which must have been interesting for them and something I would like to see more authors do.

As an academic book, this breaks the mold of abstract and depersonalized information being coldly calculated and measured and explained. Instead, it invites the reader into the experiences of a first-person narrative, with which we often identify, and then further invites us to reflect with one of the authors as they analyze the narrative. The book is itself a meta-mediational tool, using mainly stories as the mediational means to explain SCT concepts such as mediation. It is itself a ZPD constructing tool that I believe will stimulate much languaging and eventual appropriation.

Swain, Kinnear, and Steinman have produced a text whose effective scaffolding of complex knowledge might easily be modeled by many other authors. I would love to see books using this narrative approach about, for example, complexity and chaos theory and ecological linguistics. I was able to use the book this fall in two introductory graduate school classes on SCT and I found it provided useful organization to the course and accessible readings for students. I think I could easily use it with undergraduates, but only advanced level undergraduates if they are nonnative speakers of English. I have been recommending it to my colleagues at universities and to all my previous graduate students as well. My only complaint is why didn't we have such a book sooner?

References

Heath, C., & Heath, D. (2007). *Made to stick*. New York: Random House.

***The Social Psychology of English as a Global Language: Attitudes, Awareness and Identity in the Japanese Context*. Robert M. McKenzie. Dordrecht, the Netherlands: Springer, 2010. xi + 210 pp.**

Reviewed by
Christopher Starling and Yumi Tanaka
Kobe Shoin Women's University

As Kachru (1985) showed when presenting his model of world Englishes, the *outer circle*[1] and *expanding circle*[2] have historically looked to the *inner circle*[3] for guidance on English norms. Accordingly, inner circle native speakers came to be deferred to as exponents of central, norm-providing "standards," with nonstandard varieties correspondingly marginalized. More recently, with the world post-modernly de-centered, there has been increasing scrutiny of the very notion of standard, accompanied by a fresh appreciation of varieties hitherto devalued. Within the field of second language acquisition (SLA), this has meant conveying the worth of such varieties to students who might otherwise absorb inner circle myths of superiority.

It is in this context that McKenzie, in the work under review, observes the particular prestige still accorded in Japan to inner circle English and, with a view to reform, invokes social psychology to explore Japanese attitudes to and awareness of English varieties.

Organizing his book into six chapters, McKenzie devotes Chapter 1 to the worldwide spread of English and English in Japan specifically, Chapter 2 to the concept of *attitude*, and Chapter 3 to previous attitude research. An empirical study is outlined in Chapter 4 and Chapters 5 and 6 discuss its findings and their implications. Finally, appendices provide a transcription of the study material together with statistics related to follow-up analysis.

Chapter 1 begins by reviewing how English achieved its global status. The author adopts Kachru's model and then describes the place of English in Japan's education, media, and society. Chapter 2 introduces social psychology terminology and explores the concept of attitude before considering the importance of language attitudes in SLA and sociolinguistics. Chapter 3 lays theoretical groundwork for the study, detailing approaches to the measurement of language attitudes and reviewing research on attitudes of native speakers, nonnative speakers, and Japanese, specifically towards English and its varieties.

Chapter 4 describes the study, first laying out the research questions, which relate to learners' ability to identify varieties of English, their attitudes to those varieties, social variables significant in these attitudes, and attitudes towards their native language. McKenzie describes how and why he chose for evaluation speakers of Midwest and Southern U.S. English, Scottish Standard English and Glasgow Vernacular, and moderately accented and heavily accented Japanese English. He then details background variables (gender, exposure to English, regional provenance, self-perceived proficiency in English, and attitudes towards varieties of Japanese), describes his informants (558 Japanese students), and introduces his research instrument's four parts: a verbal-guise study, dialect recognition questions, a map exercise for recording perceptions of Japanese dialects, and a questionnaire concerning background variables. A pilot study is also reviewed.

In Chapter 5, McKenzie reports his findings, including that overall the inner circle speakers were ranked higher than the expanding circle (i.e., Japanese) speakers. Finally, Chapter 6 examines broader implications of the findings at considerable length in relation to the research questions posed in Chapter 4 and points the way to future research.

Interestingly, this book is a re-edition, with little change, of a doctoral thesis with a quite different title, namely *A Quantitative Study of the Attitudes of Japanese Learners Towards Varieties of English Speech: Aspects of the Sociolinguistics of English in Japan.* Arguably, this is the more appropriate title as although social psychology is important in this work, the detailed explanations relating to it are, as stated in the thesis outline (McKenzie, 2006), intended to contextualize the Japan study. One objection we might have with the later title concerns its inclusion of the word *identity.* While noting implications for identity, we found no specific treatment of identity as an issue.

That this book was written by a PhD candidate required to demonstrate relevant knowledge may have much to do with its encyclopedic approach. There are no fewer than 358 references, and the text almost constantly cites one or another of them. The same thoroughness is evident in the definition of even common terms and in the very detailed coverage of the research study.

All this erudition and care ensures a densely informative text, yet it is far from being dispassionate, instead manifesting a strong sense of mission. McKenzie's central conviction is that Japan exhibits the same "native speaker ideology" that social psychology has identified as a universal phenomenon (p. 144), but in a particularly marked form. In Japan, he believes, inner circle English (mostly in its British or American standard forms) has been

endowed with a persuasive aura of prestige and correctness. He advocates that provided intelligibility remains, students might be exposed "to as wide a range of (native and nonnative) English speech varieties as possible" (p. 161). In this position, there is evident fair-mindedness. If standards are ideological constructs, McKenzie may be seen as speaking for truth and justice in exposing and deposing them. And for sure, there is much evidence, duly cited in Chapter 1, to support his contentions that standards *are* ideological constructs and that Japan *has* hitherto privileged inner circle English and favored stereotypes of native English speakers.

The main objective of McKenzie's survey corresponds to these perceived realities. He resolves to determine "to what extent English language learners in Japan consider nonstandard or regional varieties of inner circle varieties of English as acceptable models for learning" (p. 68). For this investigation, he chooses a quantitative approach, which he sees as having multiple advantages over qualitative studies, mainly through the potential for statistical analysis and a reduction in researcher subjectivity. Significantly, his approach bypasses informant subjectivity through indirect investigation of covert attitudes that subtend variety preference. Degree of preference is assessed according to two dimensions now cardinal in social psychology: *competence* and *social attractiveness* (where the former term, often used interchangeably with *prestige* and *status*, has little to do with the *communicative competence* of SLA). Consistent with earlier findings, McKenzie's survey sees that informants deem inner circle speakers as having greater competence (status, prestige), seemingly demonstrating the strength of a native speaker ideology in Japan. Meanwhile, nonstandard inner circle speakers were held to be more socially attractive and achieved higher scores overall, implying that learners may accept them as alternative models.

Coherent though all the foregoing may seem, it can be questioned on two counts. First, with outer circle English excluded, the survey evidence appears insufficient to conclude that expanding circle speakers consider inner circle English the most competent. Second, the exclusive focus on attitudes to explain variety preference, together with the envisioning of reform on the basis of attitude findings alone, leaves no evident place for alternative, mindful criteria for variety choice. It is true that McKenzie refers to instrumental motivation in discussing other researchers, but there is no evident place for it in his own schema. Indeed, significantly, learner agency is mentioned only briefly in his book, almost as an afterthought. Focusing instead on the subconscious, McKenzie laudably aims to bypass falsehood and reveal factors the individual may not perceive or admit, but this also precludes statements

of truth. As a result, his characterization of learners tends toward caricature. With learner autonomy abstracted, any preference for inner circle varieties, and the dominance of these in the education system, can be attributed systematically to ignorance, mystification, and the assimilation of stereotypes. This is evident, for example, in the treatment of Japanese female learners, whose preference for native forms is tightly ascribed to the native speaker ideology, leading McKenzie to argue somewhat cavalierly that these women should be specifically "targeted" for re-education (p. 155). In short, for all the objectivity McKenzie accords his approach, it in fact tendentiously opens the way to a confirmation of a simplistic view. His survey can only give an ideological account of variety preference. In this regard, his project contrasts with recent efforts by others (e.g., Dörnyei & Ushioda, 2009) to express learner motivation anew in cognitive constructs.

If McKenzie oversimplifies his case theoretically, he also does little more than skirt around the practicalities of implementing the reforms he desires. For example, even if his advocacy of a range of English speech varieties in the classroom were to gain traction (which is itself far from sure), it would open up thorny issues of teacher capabilities, materials choices, and especially study time limitations.

Despite our reservations, we admire McKenzie when he remains modestly close to his implicit exigency of fairness. His most valuable message is that there remains an enduring native speaker ideology that can deeply and prejudicially influence students' attitudes to English varieties. Teachers who heed this message will beware of transmitting that ideology and instead inform students of how current standard forms achieved their primacy through political and economic power.

All in all, this book instructs both deliberately and despite itself. It is a work of meticulous, exhaustive research, lucidly written, vastly informative, and with a valid stated objective: analyzing Japanese learners' attitudes to English varieties in order to help provide a "methodological framework for the study of ideological forces" (p. 21) operating in English-learning communities, as well as to develop recognition of peripheral varieties. In this, we see much to commend. All the more regrettably, the primary focus on covert attitudes, apt as it may be in a book on social psychology, becomes controlling, and both enables and emboldens a confirmation bias that diminishes the enterprise. "Ideological forces" are normally associated with the dominant center, but this work shows how the periphery likewise develops rhetoric in its own interest.

Notes

1. Comprising ex-colonies of Britain and the United States.
2. Comprising countries where English has had no historical administrative role.
3. Comprising countries where most people are native English speakers.

References

Dörnyei, Z., & Ushioda, E. (Eds.). (2009). *Motivation, language identity and the L2 self*. Bristol, UK: Multilingual Matters.

Kachru, B. (1985). Standards codification and sociolinguistic realism. In R. Quirk & H. G. Widdowson (Eds.), *English in the world* (pp. 11-30). Cambridge: Cambridge University Press.

McKenzie, R. M. (2006). *A quantitative study of the attitudes of Japanese learners towards varieties of English speech: Aspects of the sociolinguistics of English in Japan* (Doctoral dissertation, The University of Edinburgh). Retrieved from http://www.era.lib.ed.ac.uk/bitstream/1842/1519/5/McKenzie_thesis07.pdf

Information for Contributors

All submissions must conform to *JALT Journal* Editorial Policy and Guidelines.

Editorial Policy

JALT Journal, the refereed research journal of the Japan Association for Language Teaching (*Zenkoku Gogaku Kyoiku Gakkai*), invites empirical and theoretical research articles and research reports on second and foreign language teaching and learning in Japanese and Asian contexts. Submissions from Asian and other international contexts are accepted if applicable to language teaching in Japan. Areas of particular interest include but are not limited to the following:

1. Curriculum design and teaching methods
2. Classroom-centered research
3. Cross-cultural studies
4. Testing and evaluation
5. Teacher training
6. Language learning and acquisition
7. Overviews of research and practice in related fields

The editors encourage submissions in five categories: (a) full-length articles, (b) short research reports (*Research Forum*), (c) essays on language education framed in theory and supported by argumentation which may include either primary or secondary data (*Perspectives*), (d) comments on previously published *JALT Journal* articles (*Point to Point*), and (e) book and media reviews (*Reviews*). Articles should be written for a general audience of language educators; therefore, statistical techniques and specialized terms must be clearly explained.

Guidelines

Style

JALT Journal follows the *Publication Manual of the American Psychological Association*, 6th edition (available from APA Order Department, P.O. Box 2710, Hyattsville, MD 20784, USA; by email: <order@apa.org>; from the website: <www.apa.org/books.ordering.html>). Consult recent copies of *JALT Journal* or *TESOL Quarterly* for examples of documentation and references. A downloadable copy of the *JALT Journal* style sheet is also available on our website at <www.jalt-publications.org/jj/>.

Format

Full-length articles must not be more than 20 pages in length (6,000 words), including references, notes, tables, and figures. *Research Forum* submissions should not be more than 10 pages in length. *Perspectives* submissions should be not more than 15 pages in length. *Point to Point* comments on previously published articles should not be more than 675 words in length, and *Reviews* should generally range from 500 to 1000 words. All submissions must be word processed in A4 or 8.5 x 11" format with line spacing set at 1.5 lines. **For refereed submissions, names and identifying references should appear only on the cover sheet**. Authors are responsible for the accuracy of references and reference citations.

Submission Procedure

Please submit the following materials, except for reviews, as an email attachment in MS Word format to the appropriate editor indicated below:

1. Cover sheet with the title and author name(s).
2. One (1) copy of the manuscript, with no reference to the author. Do not use running heads.
3. Contact information sheet, including one author's full address and, where available, a fax number.
4. Abstract (no more than 150 words).
5. Japanese translation of the title and abstract, if possible (no more than 400ji).
6. Biographical sketch(es) (no more than 50 words each).

Submissions will be acknowledged within 1 month of their receipt. All manuscripts are first reviewed by the Editor to ensure they comply with *JALT Journal* Guidelines. Those considered for publication are subject to blind review by at least two readers, with special attention given to (1) compliance with *JALT Journal* Editorial Policy, (2) the significance and originality of the submission, and (3) the use of appropriate research design and methodology. Evaluation is usually completed

within 3 months. Each contributing author of published articles and Book Reviews will receive one complimentary copy of the *Journal* and a PDF of the article (Book Reviews are compiled together as one PDF). *JALT Journal* does not provide off-prints. Contributing authors have the option of ordering further copies of *JALT Journal* (contact JALT Central Office for price details).

Restrictions

Papers submitted to *JALT Journal* must not have been previously published, nor should they be under consideration for publication elsewhere. *JALT Journal* has First World Publication Rights, as defined by International Copyright Conventions, for all manuscripts published. If accepted, the editors reserve the right to edit all copy for length, style, and clarity without prior notification to authors.

Full-Length Articles, Research Forum, Perspectives, and Point to Point Submissions

Please send submissions in these categories or general inquiries to:

jj-editor@jalt-publications.org

Darren Lingley, *JALT Journal* Editor

Japanese-Language Manuscripts

JALT Journal welcomes Japanese-language manuscripts on second/foreign language teaching and learning as well as Japanese-language reviews of publications. Submissions must conform to the Editorial Policy and Guidelines given above. Authors must provide a detailed abstract in English, 500 to 750 words in length, for full-length manuscripts and a 100-word abstract for reviews. Refer to the Japanese-Language Guidelines for details. Please send Japanese-language manuscripts to:

jj-editorj@jalt-publications.org

Ken Urano, *JALT Journal* Japanese-Language Editor

Reviews

The editors invite reviews of books and other relevant publications in the field of language education. A list of publications that have been sent to JALT for review is published bimonthly in *The Language Teacher*. Review authors receive one copy of the *Journal*. Please send submissions, queries, or requests for books, materials, and review guidelines to:

jj-reviews @jalt-publications.org

Greg Rouault, *JALT Journal* Reviews Editor

Address for Inquiries about Subscriptions, Ordering *JALT Journal*, or Advertising

JALT Central Office
Urban Edge Building 5F
1-37-9 Taito, Taito-ku, Tokyo 110-0016, Japan
Tel.: 03-3837-1630; Fax: 03-3837-1631
(From overseas: Tel.: 81-3-3837-1630; Fax: 81-3-3837-1631)
Email: jco@jalt.org URL: www.jalt.org

日本語論文投稿要領

*JALT Journal*では日本語で執筆された論文、研究報告、実践報告、書評等を募集しています。文体:一般的な学術論文のスタイルを用い、章立ての仕方や参考文献のデータの書き方などは、*Publication Manual of the American Psychological Association* (6th ed.)の定める方式に合わせて下さい。不明の場合は、*JALT Journal* の英語論文を参考にするか、日本語編集者までお問い合わせ下さい。なお、*JALT Journal*の読者は現場の教師が主なので、特殊な専門用語や統計的手法は、わかりやすく定義するか説明を加えるなどして下さい。原稿: 長さは、参考文献リストも含め18,000字 (書評の場合は1,500字) 以内です。A4の用紙に横書きで、1行40字、1ページ30行で印刷して下さい。手書きの原稿は受け付けません。

提出するもの:
以下の原稿を電子メールの添付書類、あるいは郵送でお送りください。

- 執筆者の名前と所属機関名を書いた表紙
- MS-Word で保存した本文(執筆者は無記名のこと)
- 執筆者連絡先(住所、電話番号、ファックス、e-mail アドレス)
- 400字以内の和文要旨
- 英文のタイトルと、500～750語の英文要旨(書評の場合は100語程度の英文要旨)
- 100字以内の執筆者略歴
- 審査を経て掲載の認められた草稿は、図表などを全て写植版にしたものにして提出すること

査読: 編集委員会で投稿要領に合っているかどうかを確認したあと、少なくとも二人の査読者が査読を行います。査読者には執筆者の名前は知らされません。査読の過程では特に、原稿が*JALT Journal*の目的に合っているか、言語教育にとって意味があるか、独創性はあるか、研究計画や方法論は適切か等が判定されます。査読は通常二か月以内に終了しますが、特に投稿の多い場合などは審査にそれ以上の時間がかかることがあります。

注意: *JALT Journal*に投稿する原稿は、すでに出版されているものや他の学術雑誌に投稿中のものは避けて下さい。*JALT Journal*は、そこに掲載されるすべての論文に関して国際著作権協定による世界初出版権を持ちます。なお、お送りいただいた原稿は返却しませんので、控を保存して下さい。

投稿原稿送り先またはお問い合わせ:

〒062-8605 北海学園大学　経営学部
JALT Journal 日本語編集者　浦野　研
電話:(011)841-1161(代) Fax:(011)824-7729
jj-editorj@jalt-publications.org

JALT Journal 第 34 巻　第 1 号

2012年4月20日　印刷
2012年5月 1日　発行
編集人　ダレン・リングリィ
発行人　ケビン・クレアリー
発行所　全国語学教育学会事務局
〒110-0016　東京都台東区台東1-37-9 アーバンエッジビル5F
TEL(03) 3837-1630; FAX(03) 3837-1631
印刷所　コーシンシャ株式会社
〒530-0043　大阪市北区天満1-18-4天満ファーストビル301 TEL(06)6351-8795