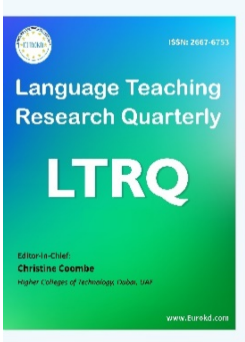

ISSN 2667-6753
Language Teaching Research Quarterly
LTRQ
Editor-in-Chief:
Christine Coombe
Higher Colleges of Technology, Dubai, UAE
www.Eurokd.com

Language Teaching
Research Quarterly

2023, Vol. 37, 125–143

Shiken Across 23 Years: James Dean Brown's Statistical Advice

Thom Hudson

University of Hawai'i at Manoa, USA

Received *02 June 2023* ***Accepted*** *24 September 2023*

Abstract

J. D. Brown (JD) served as the author of an instructional column titled "Statistics Corner" in the *Shiken: JALT Testing & Evaluation SIG Newsletter* for an extensive period spanning over two decades, from 1997 to 2019. This publication was under the auspices of the Testing and Evaluation Special Interest Group, a subdivision of the Japanese Association of Language Teaching. The audience for the columns was diverse, and included language teachers, graduate students, and language practitioners encountering real-world issues with language assessment. Throughout his tenure, JD addressed over 40 reader-submitted inquiries related to testing and quantitative research. The subjects explored in these columns can be broadly classified into two thematic areas: Second Language Testing and Second Language Research. This present article aims to provide an exhaustive examination of the diverse range of topics covered within each of these thematic categories and to trace the evolution of these subjects over the course of JD's twenty-year involvement with the newsletter. Its goal is to help situate JD's abiding concern with connecting theory and practice.

Keywords: *James Dean Brown, Language Testing and Research, Research Advice*

Introduction

JD Brown's 2005a text *Testing in Language Programs* comprehensively covers the fundamentals of language test construction, evaluation, and use. I have used the book as a class text many times, and have found that certain of JD's essential principles become clear through the text: 1) distributions underlie everything; 2) observed statistics are for scores of a particular group of examinees on a set of items under a specific set of conditions; 3) reliability is situated in a particular set of scores; 4) tests exist because someone needs to make some decision; 5) decisions always have consequences for people and institutions; and 6) validity is a characteristic of test score use, not a characteristic of a test. The corollary is: creating a test is

* Corresponding author.
E-mail address: thomhudson@gmail.com
https://doi.org/10.32038/ltrq.2023.37.06

a serious and difficult task that the test developer must work at very hard to get right. A journal issue reflecting his accomplishments should keep these principles central.

When Hassan Mohebbi, the co-editor of this special issue of *Language Teaching Research Quarterly* in honor of James Dean "JD" Brown, invited me to contribute, I had to think long and hard about what to write. I have known JD both personally and professionally for over four decades, since we were both green MA students in the TESOL program at UCLA in the 1970s. However, I did not want to write a typical tribute piece or account of a lifetime of achievement, penned by one senior scholar for another. There is certainly much room in this volume for those accounts, however, I wanted to write something that reflected the range of his academic interests and his dedication to students and the applied linguistics community while maintaining an appropriate degree of objectivity about a colleague I have worked with extensively.

As I thought about how to focus my contribution, I recalled how JD was always busy in some activity related to his work. He is a prolific researcher and author, involved in many different aspects of our profession. JD has served on over 100 MA and PhD thesis and dissertation committees, authored numerous chapters, monographs, books, and research articles, and served on editorial boards and professional committees. He regularly delivered conference presentations, colloquia, and workshops, advised students and colleagues on their research, and conducted program evaluations both domestically and internationally. These endeavors all related to his interests in language assessment, applied linguistics, and language pedagogy more broadly.

In the midst of all these activities, one particular endeavor that I believe exemplifies JD's wide-ranging engagement in the field was his involvement with the column *Statistics Corner*. For over twenty years, from 1997 to 2019, he served as the author for this column in the *Shiken*: *JALT Testing & Evaluation SIG Newsletter*, which is produced by the Testing and Evaluation Special Interest Group of the Japanese Association of Language Teaching. In my contribution here, I will reflect on the diverse range of academic topics that he covered throughout the years in this column. I have always appreciated JD's commitment to answering the questions submitted to the column, despite having numerous other professional commitments. However, before delving into the specific topics covered in *Statistics Corner*, some background and context are necessary.

Shiken Statistics Corner: Origins

The *Shiken* column that JD regularly contributed was envisioned to fill a particular niche for the foreign language teaching community in Japan beginning in the late 1990s. Foreign language education is populated with multiple constituencies with questions about how to successfully carry out their professional and academic work. The constituents represent teachers, researchers, administrators, and curriculum designers. All these groups at various times confront the need to deal with language assessment and address instances of their own professional lack of exposure and uncertainty. The *Shiken* column was envisioned as a venue to help the community deal with some of the challenges presented by testing and research.

JD Brown has had a long involvement with the foreign language education setting in Japan. He first visited Japan in 1987 when he taught a short course for Temple University Japan (TUJ) (Brown, 2016b). This was the beginning of a long-term involvement of teaching courses and workshops, student advising, and conference presentations throughout Japan. Because of his disciplinary focus, his topics often centered around language assessment and methods of test data analysis. Most people who become engaged with the foreign language educational system in Japan encounter questions around university entrance examinations at some point. This generally leads to other discussions around examinations in Japan more generally, which leads to even more discussions about assessment. JD was no exception to this unravelling thread. He, along with colleagues, contributed articles raising pertinent issues about ways in which testing had been envisioned and instituted (Brown, 1987, 1990, 1995; Brown & Yamashita, 1995).

In 1995, he was invited to give a plenary speech at the JALT Conference in Nagoya. This invitation specifically requested that he address English language entrance examinations in Japan. At this conference, the first meeting of the JALT Testing and Evaluation (TEVAL) special interest group (SIG) was held. The formation of the SIG led to the establishment of the newsletter, *Shiken: JALT Testing & Evaluation SIG Newsletter*.

The first *Statistics Corner: Questions and answers about language testing statistics* contribution by JD appeared in 1997, Volume 1 Number 1 of *Shiken*. Between 1997 and 2019, he produced some 50 columns following a common question-and-answer format. Questions were submitted to him from the general language community of teachers or graduate students, in conversations with colleagues, newsletter editors, or other readers from Japan, Hawai'i or elsewhere who were curious about some aspect of language assessment, data analysis, or research. JD notes that, "Regardless of their source, all of the questions were raised by people other than me who were interested in testing, research, or statistics... (Brown (2016a, xv)".

In approximately 2016, TEVAL officers suggested that JD's contributions to *Shiken* be compiled into an edited book, which was published the same year (Brown, 2016a). In this book, JD organized the columns into related topics rather than presenting them in chronological order. The topics roughly fall into two categories: Second Language Testing and Second Language Research (Brown, 2016a). The Language Testing category includes columns on Testing Strategies, Item Analysis, and Reliability Issues while the Second Language Research category includes columns on Planning Research, Interpreting Research, and Research Analyses. It is important to note that these categories were not predetermined when the columns were written, but were developed for the purpose of organizing the 2016 book. The rest of this article will focus on the contents of the columns within their respective topic categories[1]. The columns are presented using the overall topic organization of the 2016 book and are organized chronologically within each topic category. Table 1 below shows the structure of the *Shiken* organization in the 2016 book. Additionally, this article includes coverage of the seven columns JD submitted to *Shiken* after the book collection was published in 2016, and have been incorporated into the appropriate content categories.

[1] Brown (2001b & 2004a) were deemed to be redundant with material discussing eigenvalues and Yates' correction that was covered more comprehensively in subsequent chapters. Thus, they and are not included in the 2016 book or in this review.

Table 1

Brown Shiken Column Topics (Adapted from Brown, 2016)

Part 1: Second Language Testing				
Column Title		Year	Vol	#
Section 1: Testing Strategies				
1	University Entrance Examinations: Strategies for creating positive washback on English language teaching in Japan	2000	3	2
2	What is construct validity?	2000	4	2
3	What is two-stage testing?	2001	5	2
4	Extraneous variables and the washback effect	2002	6	2
5	Test-taker motivations	2004	8	2
6	Resources available in language testing	2006	10	1
7	Solutions to problems teachers have with classroom testing	2013	17	2
8	Differences in how norm-referenced and criterion-referenced tests are developed and validated	2014	18	1
9	Testing intercultural pragmatics ability	2015	19	1
10	Developing and using rubrics: Analytic or holistic?	2017	21	2
11	Developing rubrics: What steps are needed?	2018	22	1
12	What is assessment feedback and where can I find out more about it?	2019	23	1
13	Overall English proficiency (whatever that is).	2019	23	2
Section 2: Item Analyses				
14	How can we calculate item statistics for weighted items	2000	3	2
15	What issues affect Likert-scale questionnaire formats?	2000	4	1
16	What is a point-biserial correlation coefficient?	2001	5	3
17	Distractor efficiency analysis on a spreadsheet	2002	6	3
18	Norm-referenced item analysis (item facility and item discrimination)	2003	7	2
19	Criterion-referenced item analysis (The difference index vs. the B-index)	2003	7	3
20	Likert items and scales of measurement	2011	15	1
Section 3: Reliability Issues				
21	Reliability of surveys	1997	1	2
22	Cloze tests and optimum test length	1998	2	1
23	The standard error vs. standard error of measurement?	1999	3	1
24	Can we use the Spearman-Brown prophecy formula to defend low reliability?,	2001	4	3
25	The Cronbach alpha reliability estimate	2002	6	1
26	Generalizability and decision studies	2005	9	1
27	How do we calculate rater/coder agreement and Cohen’s kappa?	2013	16	2
28	Consistency of measurement categories and subcategories	2016	20	2
29	Calculating reliability of dictation tests: Does K-R21 work?	2018	22	2
Part 2: Second Language Research				
Section 4: Planning Research				
30	Characteristics of sound qualitative research	2005	9	2
31	Characteristics of sound quantitative research	2015	19	2

32	Characteristics of sound mixed methods research	2016	20	1
Section 5: Interpreting Research				
33	Skewness and kurtosis	1997	1	1
34	Generalizability from second language research samples	2006	10	2
35	Sample size and power	2007	11	1
36	Sample size and statistical precision	2007	11	2
37	The Bonferroni adjustment	2008	12	1
38	Effect size and eta squared	2008	12	2
39	Confidence intervals, limits, and levels?	2011	15	2
40	What do distributions, assumptions, significance vs. meaningfulness, multiple statistical tests, causality, and null results have in common?	2012	16	1
Section 6: Research Analyses				
41	The coefficient of determination	2003	7	1
42	Principal components analysis and exploratory factor analysis—Definitions, differences, and choices	2009	13	1
43	Choosing the right number of components or factors in PCA and EFA (Overlap with #10)	2009	13	2
44	Choosing the right type of rotation in PCA and EFA	2009	13	3
45	How are PCA and EFA used in language research?	2010	14	1
46	How are PCA and EFA used in language test and questionnaire development?	2010	14	2
47	Chi-square and related statistics for 2 x 2 contingency tables (Overlaps with Chapt. 19)	2013	17	1
48	Consistency in research design: Categories and subcategories	2017	21	1

As a result of space constraints, not every column will be explicitly addressed individually here. The discussion will focus primarily on the conceptual implications of JD’s work. Several of the columns he produced are technical and procedural in nature. Those columns, such as those emphasizing spreadsheet calculations, are not directly addressed in the review. Also, it should be kept in mind that since the *Shiken* columns are generally addressing specific questions, they are relatively short and concise. In recognition of this constraint, JD includes many references within the columns which the *Shiken* reader can follow to pursue the topics in greater depth.

Part 1: Second Language Testing: Testing Strategies, Item Analysis, Reliability Issues

Columns focusing on second language testing concerns address many of the questions that arise from day-to-day test use.

Testing Strategies

The first *Shiken* column, shown in Table 1, is 1 *University Entrance Examinations*. While this column was initially a stand-alone submission, not part of the Q-and-A format of the other *Statistics Corner* columns, as noted above its topic is one that was foundational in the initial establishment of the JALT Testing and Evaluation SIG. Additionally, his discussion highlights one of his recuring themes about the role of examinations: tests exist within social contexts and have effects. The column is framed with the question-and-answer format:

> *QUESTION: For many years, you have been criticizing the English entrance examinations used by Japanese universities. Has any of that taught you what kinds of positive responses might be useful for solving the problems these tests create?*
> *ANSWER: I think the best strategy that can be used at the moment to solve the university entrance examination problem would be to work to turn them into positive forces for change. Thus, this chapter explores some of the ways the university entrance examinations in Japan could be used to foster positive washback effects on English language instruction. During the last twelve years, a great deal has been written about the quality and appropriateness of examinations in Japan.*

In much of this column the focus is on the potential for positive instructional washback effects in Japanese examinations. The suggestions here are closely linked to positive curricular goals and practices, reflecting an orientation towards the examination system that views examinations as more than just context-free standardized tests, a dominant notion held not only in Japan but in many other countries as well. His concerns directly relate to the consequences of the examination system on how teachers will teach and how students will study. JD's paper emphasizes the importance of teamwork and collaboration between teachers and the university examination writers to foster positive washback. The strategies needed are organized into four categories: test design, test content, logistical strategies (local interactions), and interpretation strategies. This framework highlights the complexity of assessment issues and reflects JD's consistent viewpoint throughout this work. Concern with how tests may have unintended consequences in their washback effects is also a highlight in column 4, *Extraneous Variables and the Washback Effect*. The motivating question that was submitted has a wide scope, but its central query is:

> *Question: In your 1988 book Understanding Research in Second Language Learning you mention different types of extraneous variables such as subject expectancy, the halo effect, and the Hawthorn effect… [W]hat is the relation of these terms to washback?*

In addressing the question at hand, JD discusses the potential impact of environmental factors, sample selection, and measurement issues on the validity of a study's interpretation. These concerns reoccur in later columns. However, JD's response extends beyond these concerns to encompass the concept of washback mentioned above. By providing definitions of washback, JD highlights its relationship to the influence of testing context on teacher and learner actions that either promote or inhibit language learning. As such, washback may be positive, negative, or a mixture of the two, and be subsumed under many terms such as test impact, test feedback, measurement-driven instruction, etc. The implications of washback are such that its classification as positive or negative is contingent upon whether the test is effectively achieving its desired goals. Here is where the connection to validity becomes central in JD's argument. For example, if the educational setting has defined language learning as grammar-translation, then a test that focuses on achieving outcomes from grammar-translation instruction has, by definition, positive feedback. Conversely, if changes to the curriculum shift towards communicative, task-based, or language-for-specific-purposes oriented instruction, that same examination will likely produce negative washback.

Extending this discussion, within a framework that foregrounds test effects, there is a necessary connection to construct validity, which is the topic of Column 2, *What is construct validity?* The submitted question to which JD responded began as follows:

> *Question: Recently I came across an article mentioning that a test had poor construct validity. What exactly is construct validity?...*

JD's answer draws on fundamental definitions of construct validity, where it is defined as the experimental demonstration of a test's ability to measure the construct it claims to measure (Brown 2016a, p. 28). He notes that at its most simplistic level, such an experimental demonstration would simply involve administering the test to a group of individuals possessing the construct (Group 1) and a group who do not possess the construct (Group 2), and comparing the performance of the two groups. If Group 1 performs better than Group 2 to a non-trivial extent then the exam has shown evidence of construct validity. JD discusses several threats to interpretations of construct validity, but argues that careful and systematic planning can mitigate many of the design concerns. However, JD emphasizes that the traditional, narrow view of test validation needs to be expanded to include the implications and consequences of test results. To this end, he cites Messick's (1988) call for a unified and expanded theory of validity that considers both score interpretation and use, incorporating judgments of value implications and social implications.

The importance of congruence between test score purpose and the test development process is highlighted as a crucial factor in accounting for the validity of implications drawn from test scores. In addressing this issue of test score and test use is JD's consistent attention to the differences between norm-referenced (NRT) and criterion-referenced test (CRT) development. Column 8 of the *Shiken* contributions reflects that attention. The question that JD addressed was as follows:

> *QUESTION: What are the major differences between norm-referenced and criterion-referenced tests? How can these two tests be developed and validated? [Submitted by a participant in the Kuroshio (Aloha Friday) Seminar that Kimi Kondo-Brown and I conducted on May 23, 2014 at the Bunkyo Civic Center in Tokyo]*

In summarizing the differences between the two types of tests, it is clear that those differences hinge on the types of decisions that are going to be made based on test results. One of the functions of highlighting these differences is to raise awareness and emphasize the importance of making conscious decisions when developing and interpreting tests. The tests differ primarily in terms of:

1. How scores are to be interpreted. Do scores compare test takers (NRTs) or reflect amount of material known (CRTs)?

2. How specific the test material is. Do scores measure general language ability (NRTs) or specific skills or knowledge (CRTs)?
3. How the distribution of scores is expected to look. Are the scores designed to become normally distributed (NRTs) or is it okay to have all high scores on a post test and all low scores on a pre-test (CRTs)?

The column also addresses development and validation strategies for the two types of tests in a clear pair of tables which contrast NRT (Standardized) and CRT (Classroom) approaches. These tables provide techniques for establishing that the tests produce data for the types of decisions that will result from the test administration. In the concluding section of this column JD specifically mentions that practicing language teachers reading the column should realize that most of the testing they do in classrooms is CRT, and that it would be helpful to read the information in this column along with the previous column, Column 7 titled "Solutions to problems teachers have with classroom testing", which focuses on the classroom construction of CRTs. JD points out that teachers' test writing, development, and validation practices often suffer due to the tendency to treat tests as an afterthought. Writing tests is often considered the least satisfying part of the teaching work, and is not seen as integral to the instructional process.

Item Analysis

The second section of columns in the *Second Language Testing* section addresses test item analysis. These columns are often statistically procedural and specific to particular types of test item construction. So, the specifically procedural columns will not be treated separately in this discussion. In general, the topics address how systematic item analysis is both necessary and complicated in item evaluation. Two of the columns however, 18, Norm-referenced item analysis (item facility and item discrimination), and 19, Criterion-referenced item analysis (the difference index vs. the B-index), address how NRT and CRT item analyses are treated differently. They provide a summary of strategies (procedural steps) for developing and validating the two types of tests.

This segment of the contributions to *Shiken* focuses specifically on the test item as object. Initially, the test item is addressed in both columns with first principles. What are the steps involved? For both NRT and CRT exams, the steps are the same:

1. Assemble or write a relatively large number of items of the type you want on the test.
2. Analyze the items carefully using item format analysis to make sure the items are well written and clear.
3. Pilot the items using a group of students similar to the group that will ultimately be taking the test.
4. Analyze the results of the pilot testing using item, analysis techniques (for either NRT or CRT purposes).
5. Select the most effective items (and get rid of the ineffective items)

The way in which this part of the process is foregrounded harkens back to JD's previously noted emphasis on how the testing development process needs to be systematic. Again, tests are not to be seen as an afterthought.

The basic purpose of any NRT is to spread examinees out along a continuum in order to make decisions about aptitude, proficiency, or placement. Two statistics that help in the selection of items achieve this are item facility (IF) and item discrimination (ID). The basic purpose of a CRT is to measure the amount of relevant material that examinees know in order to make decisions about achievement or diagnosis. Two item level statistics to do this are the Difference Index (DI) and the B-Index. The differences among these four indices can be seen in Table 2.

Table 2

Item Level Statistics for Item Analysis

NRT	CRT
IF: the proportion of examinees who answer an item correctly	DI: the difference of the item facility on the pretest and the item facility on the post test
ID: the difference between the IF for the upper-level examinees and the IF for the lower-level examinees	B-Index: the item facility on the item for the examinees who passed the test minus the item facility for the examinees who failed the test

These item analysis statistics point out how comparative groupings are made in the two different families of test purpose. The NRT groupings order examinees relative to other examinees in their performance on the examination. The CRT groupings, on the other hand, order examinee results in relation to groups that either have demonstrated control of the material or in terms of whether they are expected to have gained control of the material because of presence or absence of content exposure. Other approaches to item analysis are discussed in the *Shiken* columns. Correlational analyses (e.g., point-biserial), analyses of distractor efficiency in multiple-choice exam items, how to address weighted items, and ways of addressing Likert and scale item formats are discussed.

One of the Shiken question submissions in Column 15, *What issues affect Likert-scale questionnaire formats?* and JD's response to the question, was very revealing of how JD has seen his role as guide rather than a judge. The question submitted related to some issues with Likert type items.

> *Question: Recently I came across a survey which attempted to evaluate student interest about a range of classroom topics. Students were asked to rank their interest in various potential topics according to this scale:*
>
> *10 if they felt a topic was interesting*
> *6 if they felt a topic was above average interest*
> *4 if they felt a topic was below average interest*
> *1 if they felt a topic was not worth studying in class*
>
> *Please note that only four responses were permitted: 10, 6, 4, and 1. Is this an acceptable survey design? Should the scale reflect the number of permissible responses rather than an arbitrary figure of 10?*

There are clearly some problems with the scale that is the focus of the question. JD's response transformed the question into a lesson in the problems associated with designing such scales. The discussion offers a thorough coverage of Likert scales such as those used to elicit respondents' opinions or feelings about a topic in scale categories like 1 = very serious to 4=

not serious at all. In many cases, these scales have an implied assumption of equal distances among the categories or mutually exclusive categories. The difference between 1 and 2 is assumed to be the same as the difference between 2 and 3, etc. Likewise, the selection of the second category, *above average interest*, prevents the simultaneous selection of the third category, *below average interest*. JD identifies the potential problems in determining whether the scale that is being used is categorical, rank-ordered, or continuous in nature.

He notes that the 1, 4, 6, 10 scale that is the focus of *Shiken* column 15 is a "strange scale indeed" (97). The scale is not really continuous since the points on the scale are not equal interval. The number of intervals differ between the scale points **10** 9 8 7 **6** 5 **4** 3 2 **1**. Similarly, the scale is not ordinal since the designations of 10th, 6th, 4th, and 1st rankings do not make sense when there are only 4 possible rankings. Likewise, the scale is not categorical since the categories do not make sense in relation to one another. Not only is the nature of the different scales unclear, but scale point 1 is asking about a very different construct from the other three scale categories. JD suggests that the scale would have been much better as one of the three distinct types of scale rather than the one here which was "neither fish nor fowl" (98). He concludes by noting that the scale 'must have been very difficult indeed to analyze and interpret" (98).

Reliability Issues

The nine columns in the *Reliability Issues* section are concerned with consistency within measurement. Consistency is a critical concern in testing, referring to the fundamental requirement of stability or reliability of test results over time or across different testing conditions. Several approaches to establishing consistency have been developed. Test-retest reliability refers to the consistency of results when the same test is administered to the same person on different occasions. Internal consistency refers to the consistency of results within a single test or measure. Inter-rater reliability refers to consistency of results when the same test is administered by different raters or examinees.

Five of the columns in this section focus primarily on procedural techniques for using reliability information in the test revision and development process: 21 *Reliability of surveys*, 22 *Cloze tests and optimum test length*, 23 *Can we use the Spearman-Brown prophecy formula to defend low reliability?* 26 *Generalizability and decision studies*, and, 29 *Calculating reliability of dictation tests: Does K-R21 work?*. These columns emphasize the goal of maximizing consistency within the practical constraints of the testing context. These columns demonstrate the role the *Shiken* columns played in answering specific local questions by the readers.

More generally, three of the *Shiken* columns serve to highlight JD's specific concern with the differences between concepts of consistency between NRT and CRT contexts: 25 *The Cronbach alpha reliability estimate*, 27 *How do we calculate rater/coder agreement and Cohen's kappa*, and 28 *Consistency of measurement categories and subcategories*. NRT approaches focus on how consistently the test establishes the examinee's relative standing among other examinees while CRT focuses on how consistently the examinees are given a particular classification. This distinction is captured in JD's discussion of column 25, 27, and 28 by focusing on how consistency is classified. The Cronbach alpha coefficient estimates the amount of variance that is systematic within a set of scores and indicates the extent to which all of the test items are behaving in the same way across the examinees. JD points out that Cronbach alpha applies to NRT scores and decisions, and works best with data that are normally distributed. CRTs are evaluated according to consistency of the decisions made based on the test scores. The types of consistency that are considered in evaluating the test relate to how the scores consistently classify the examinees. The agreement coefficient and Cohen's kappa coefficient work within this decision framework.

Throughout his discussion, JD works to show the readers of the columns how the type of consistency that they need to attend to depends upon the uses of the examination results. He shows that there are technical requirements for both types of tests, and that neither test is necessarily good nor bad. However, the particular uses might be appropriate or inappropriate.

Part 2: Second Language Research: Planning Research: Planning Research, Interpreting Research, Research Analyses

Columns addressing Second Language Research in the second part of the 2016 book transition from the previously discussed strategies of language testing *per se* to issues that focus more on measurement approaches within second language research more broadly. Specifically, the columns in this second section address the three areas of *planning research, interpreting research*, and research *analyses*.

Planning Research

The columns under the *Planning Research* section (30, 31, & 32) address sound qualitative, quantitative, and mixed-methods research in order. For context, it is noted here that the first column, addressing qualitative research, appeared as a *Shiken* column in 2005. The second two columns, addressing quantitative and mixed-methods research appeared in 2015 and 2016, responding to the reader question for sound quantitative research below:

> *QUESTION: In [the Brown 2005 column] you explained the characteristics of well-done qualitative research by explaining the importance of dependability, credibility, confirmability, and transferability. You mentioned in passing that the parallel characteristics for quantitative research were reliability, validity, replicability, and generalizability. But you never really explained those quantitative research characteristics. I think it would be useful to know more about these characteristics of sound quantitative research and maybe even something about the characteristics of good quality mixed-methods research. Could you talk about these other research paradigms?*

In addressing the application of different research paradigms for his readers, JD defines research in the first of the columns, 30 *Characteristics of sound qualitative research*, as "any systematic and principled inquiry" (2005b, p. 31). Across the different columns in this section, he posits that "sound qualitative research (at one end of the continuum) can be systematic in terms of its dependability, credibility, confirmability, and transferability, while sound quantitative research can be systematic in terms of its reliability, validity, replicability, and generalizability" (31, p. 161). For mixed-methods research, he notes that mixed methods must address all these areas of systematicity. However, he makes a stronger claim as well with reference to mixed-methods research. He argues that mixed methods is not merely a collection of the two methodologies, but rather a method which combines the two research paradigms in a systematic and principled way. This systematic and principled argument attempts to show the *legitimacy* of how the different methods support each other.

The essential advice that JD is giving the readers of his *Shiken* columns is that dogmatic adherence to a particular paradigm may not be the most productive approach to their own research or to their critiques of the research of others. Rather, it is important to accept that evidence can come from different approaches. What is key is the *systematic* analysis of the research in terms of its systematicity and adherence to sound data analysis, whether the data are quantitative, qualitative, or mixed.

Interpreting Research

This section, *Interpreting Research*, presents some of the trickiest issues in understanding reported research results as well as with grasping the consequences of research design choices. The scope of the section is reflected in the question-and-answer introduction to Column 40 (Brown, 2012a, p. 27).

> *QUESTION: The field of statistics and research design seems so complicated with different assumptions, and problems associated with each form of analysis. Is there anything simple? I mean are there any principles that are worth knowing that apply across the board to many types of statistical analyses?*
> *ANSWER: Fortunately, a number of issues are common to the most frequently reported forms of statistical analysis. In this chapter, I will discuss a number of those issues in the following six categories: distributions underlie everything else, assumptions must be examined, statistical significance does not assure meaningfulness, multiple statistical tests cloud interpretations, causal interpretations are risky, and null results do not mean sameness.*

This question, and the columns that follow, address statistical challenges in traditional quantitative analyses based on normal distributions. Analyses based on assumptions of normal distributions are the most frequent statistical analyses used in language studies. JD, in the "ANSWER" section, emphasizes the foundational message throughout the Shiken columns that "distributions underlie everything else". In Column 33, *Skewness and Kurtosis* (Brown, 1997a), JD discusses the nature of normality in distributions and how the data may deviate from normality due to such factors as the data being collected from either a test prior to instruction or post instruction. His discussion emphasizes the need to consider both the symmetry and the spread of the scores.

Another concern throughout these columns focuses on the problems with multiple statistical tests within quantitative studies. The phenomenon of multiple statistical tests occurs when researchers carry out multiple statistical comparisons without adjusting the probability level for rejecting a null hypothesis. Each experimental comparison is required to set a statistical level of possibility that the inference may be incorrect. This is known as the *alpha* level. Basically, each statistical test that is carried out on samples has a possibility of leading to an incorrect conclusion that there is a statistical difference within populations. When multiple tests are carried out, each one has its own chance that the inference is erroneous. There is a natural inclination for novice researchers to try to answer as many possible questions as they possibly can with the data that they have. However, the more statistical tests that are carried out independently on the same data, the greater the likelihood that one of the inferences will be false. The problem is that we do not know which one(s) might be wrong. Brown directly addresses this in *Shiken* column (37) on the Bonferroni Adjustment (Brown, 2008a). The Bonferroni adjustment is a statistical technique that spreads the *alpha* level across all comparisons in the experiment. While this seems like a win-win situation, the approach requires that each comparison meet a more conservative significance level in order to keep the overall experiment-wide *alpha* level at the acceptable level. With several tests, the significance level for each test can become so restrictive that virtually no differences can be found to be statistically significant.

JD continues his concerns with distributions and statistical decisions in columns that emphasize the role of sample size in research design, 34 *Generalizability from second language research samp*les, 35 *Sample size and power*, and 36 *Sample size and statistical precision*) (Brown 2006b, 2007a, 2007b). These contributions discuss the need for researchers to ensure adequate sample sizes in order to account for variations in distributions, actual strength of effects, and soundness of measurement. It is important to have sufficient numbers of research participants to get generalizable results. If the sample size is too small, it can negatively affect the interpretation of the results. This is because of the role played by something called the null hypothesis (H_0), which posits that there is no difference between the groups being studied. Because of this initial hypothesis, it is up to the researcher to prove that there is a difference. JD quotes Fisher (1971) who wrote that the null hypothesis is “the hypothesis that the phenomenon to be demonstrated is in fact absent (p.13). With larger sample sizes, it becomes easier to prove that there is a difference when there is one. In other words, larger sample sizes increase the “power” of the study, making it more likely that significant differences will be detected. So, it is important for researchers to use an appropriate sample size in order to accurately demonstrate any differences between the groups being studied.

JD also discusses “effect size” in column 38, *Effect size and eta squared* (Brown, 2008b). Effect size measures how strong the explanatory value of the targeted variable(s) in the study is. Effect size allows for a stronger claim than merely stating that the observed effects are greater than chance and provides an indication of the strength of any experimental intervention. This, then, allows a claim stronger than merely that the observed effects are greater than chance. It is an indication of whether any experimental intervention was strong. This concern is congruent with his caveat (Brown, 2012a) that statistical significance does not assure meaningfulness. A result can be non-random, but still unimportant. JD advises the *Shiken* readers to aim for a more impressive statement than, “my variable is stronger than nothing at all.”

Research Analyses

The final topic division of the columns in Table 1 is designated as *Research Analyses*. It includes discussions of the coefficient of determination (which indicates how to interpret the values of a correlation coefficient), principal components and factor analysis, and consistency in research.

However, most of this section is comprised of the five columns focusing on various issues with PCA (principal components analysis) and EFA (exploratory factor analysis). These columns comprise all of JD's columns for 2009-2010, *Shiken* volumes 13 and 14. As such, they represent a bit of a digression from the patterns of column topics throughout the 1997-2019 period in which the columns appeared. Given the divergence in structure involved with these columns from the normal *Statistics Corner* contents, I contacted JD to find out how they had developed. He responded in an email as follows:

> I had been teaching EFA/PCA repeatedly for a number of years at [Temple University Japan] as part of the advanced stat course to doctoral students in [Tokyo] and Osaka. The questions they asked in class served as the basis for various handouts that I had tracked down and created for explaining the underlying practical answers. …Given the timing of the columns you're referring to, however, my guess is that the catalyst for these columns was questions that were raised by the MA (and PhD) students taking not only [the stats course], but also the survey research course (mostly MA Ss). No way to teach survey research to MA Ss without EFA/PCA, but I had to keep it really clear and simple. … So, by the end of that whole process over many years, I had answers to the students' central, practical questions and I put them in those columns. The bottom line: the need to explain and address varying levels of students' questions over and over and over and over in increasingly simpler and clearer terms resulted in these columns… (4/18/2023).

So, these *Statistics Corner* columns were created to address questions particular to a group of the *Shiken* audience – graduate students who were conducting or reading advance-level research involving multiple content factors contributing to dependent variables. These questions had been consistently raised by the students he had been teaching in MA and PhD level classes at TUJ. As a result, he organized his answers to these questions systematically into the five columns.

An explanation for the specific difference in focus and orientation of these columns from the others can be glimpsed from the question posed at the beginning of the first column, 42 *Principal components analysis and exploratory factor analysis: Definitions, differences, and choices* (Brown, 2009a).

> *QUESTION: In Chapter 7 of the 2008 book on heritage language learning that you co-edited with Kimi Kondo-Brown, there is a study (Lee and Kim, 2008) comparing the attitudes of 111 Korean heritage language learners. On page 167 of that book, a principal components analysis (with varimax rotation) describes the relationships among 16 purported reasons for studying Korean with four broader factors. Several questions come to mind. What is a principal component analysis? How does principal components differ from factor analysis? What guidelines do researchers need to bear in mind when selecting "factors"? And, finally, what is a varimax rotation, and why is it applied?*

Throughout the development of the PCA/EFA thread, the following five questions emerged as the basis for the five *Shiken* columns:

1. What are principal components analysis (PCA) and exploratory factor analysis (EFA)?
2. How do investigators determine the number of components or factors to include in the analysis?
3. What is rotation, what are the different types, and how do researchers decide which particular type of rotation to use?
4. How are PCA and EFA used in language research?
5. How are PCA and EFA used in language test and questionnaire development?

In the first of these columns, 42, JD showed what PCA and EFA were, and to some extent how they should be presented and interpreted. He also defined basic notions of factor loadings, communalities, and the mathematical basis for PCA and EFA. In the second column 43, *Choosing the right number of components or factors in PCA and EFA* (Brown, 2009b), he addressed several of the methods for determining how to decide how many factors should be included in the model. Some of the rules incorporate theory-based numbers of factors while others are exploratory in nature. Column 44, *Choosing the right type of rotation in PCA and EFA* (Brown, 2009c) presented definitions of *rotation.* These are mathematical methods used to align factors to theoretical entities. Some of these are applied differently depending upon whether the underlying theory assumes that the factors are correlated with one another or are assumed to be uncorrelated. JD suggests it is often useful to try different methods to decide which works best. Column 45, *How are PCA and EFA used in language research* (Brown, 2010a) suggests how EFA and PCA can be used in language research to reduce the number of variables in a study, explore patterns in intercorrelations, and support theory. The final of the columns on PCA/EFA, 46, *How are PCA and EFA used in language test and questionnaire development?* (Brown 2010b), expands the discussion to argue specifically for the use in developing tests and questionnaires. Here, he is essentially demonstrating how PCA/EFA can be used as useful tools in the creation and development of assessment instruments.

Thus, these five columns work together in a systematic manner to guide the *Shiken* readers through a complex analytic procedure focusing on topics specific to their research interests.

Conclusions: Context in Interpretations; Effects of Testing; Interpretation of Significance Perspective on Findings

The topics covered over the 23-year tenure that JD helmed the *Statistics Corner* show remarkable scope and exhibit a keen awareness of the varying levels of knowledge among the column's readers. Throughout the columns, JD emphasized the importance of context and purpose in interpreting statistics and making decisions based on language tests.

A central concern throughout the columns his acknowledgement that language tests are used for some purpose in order to make some decision. Context is important. Consequently, throughout his columns, there is a necessary and appropriate reluctance to provide un-nuanced answers addressing what is "good," "right", "desirable" which are not closely linked to assessment purpose and test use. Indeed, in the very first *Shiken* column in 1997 JD states

> "One last point I would like to make: the skewness and kurtosis statistics, like all the descriptive statistics, are designed to help us think about the distributions of scores that

> our tests create. Unfortunately, I can give you no hard-and-fast rules about these or any other descriptive statistics because interpreting them depends heavily on the type and purpose of the test being analyzed. Nonetheless, I have tried to provide some basic guidelines here that I hope will serve you well in interpreting the skewness and kurtosis statistics when you encounter them in analyzing your tests. But, please keep in mind that all statistics must be interpreted in terms of the types and purposes of your tests." (Brown, 1997a, p. 23)

The *Shiken* columns have provided valuable insights into the complexities of language learning research. The articles consistently cautioned against oversimplification by emphasizing the importance of avoiding prescriptive absolute statements and by acknowledging the various approaches and assumptions used in both NRT and CRT. In addition, JD discussed the importance of using both quantitative and qualitative research methods to better understand the nuances of language learning. Furthermore, the articles highlighted the critical role of assessment in pedagogy and its potential washback effects on language learners.

One critical point emphasized in the *Shiken* columns was the need to recognize that assessment is not be seen as an afterthought, as something to be added as punctuation to the serious work of education. Assessment has consequences for students, teachers, and educational institutions.

Overall, JD's *Shiken* columns offer a wealth of information and insights into the intricacies of language testing and research. They serve as a valuable resource for anyone interested in delving deeper into the complexities of language learning research, and provide a nuanced perspective on the challenges faced by researchers in this field. The articles underscore the fact that there are no simple solutions to the multifaceted issues facing researchers in the field of language learning, though we may wish that there were.

ORCID

https://orcid.org/0009-0005-1102-4141

Acknowledgements

Not applicable.

Funding

Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

Rights and Permissions

Open Access

References

Brown, J. D. (1987). False beginners and false starters: How can we identify them? *The Language Teacher*, 11(4), 9-11.

Brown, J. D. (1990). Where do tests fit into language programs? *JALT Journal*, 12(1), 121-140.

Brown, J.D. (1995). English language entrance examinations in Japan: Myths and facts. *The Language Teacher,* 19(10), 21-26.

Brown, J. D. (1997a). Statistics Corner: Questions and answers about language testing statistics: Skewness and kurtosis. *Shiken: JALT Testing & Evaluation SIG Newsletter, 1*(1), 16-18.

Brown, J. D. (1997b). Statistics Corner: Questions and answers about language testing statistics: Reliability of surveys. *Shiken: JALT Testing & Evaluation SIG Newsletter, 1*(2), 17-19.

Brown, J. D. (1998a). Statistics Corner: Questions and answers about language testing statistics: Cloze tests and optimum test length. *Shiken: JALT Testing & Evaluation SIG Newsletter, 2*(2), 19-22.

Brown, J. D. (1999a). Statistics Corner. Questions and answers about language testing statistics: The standard error vs. standard error of measurement? *Shiken: JALT Testing & Evaluation SIG Newsletter, 3*(1), 15-19.

Brown, J. D. (2000a). University Entrance Examinations: Strategies for creating positive washback on English language teaching in Japan. *Shiken: JALT Testing & Evaluation SIG Newsletter, 3(2),* 4-8.

Brown, J. D. (2000b). Statistics Corner. Questions and answers about language testing statistics (How can we calculate item statistics for weighted items?). *Shiken: JALT Testing & Evaluation SIG Newsletter, 3(2),* 19-21.

Brown, J. D. (2000c). Statistics Corner. Questions and answers about language testing statistics: What issues affect Likert-scale questionnaire formats? *Shiken: JALT Testing & Evaluation SIG Newsletter, 4(1),* 18-21.

Brown, J. D. (2000d). Statistics Corner. Questions and answers about language testing statistics: What is construct validity? *Shiken: JALT Testing & Evaluation SIG Newsletter, 4(2),* 7-10.

Brown, J. D. (2001a). Statistics Corner. Questions and answers about language testing statistics: Can we use the Spearman-Brown prophecy formula to defend low reliability? *Shiken: JALT Testing & Evaluation SIG Newsletter, 4(3),* 7-9.

Brown, J. D. (2001b). Statistics Corner. Questions and answers about language testing statistics: What is an eigenvalue? *Shiken: JALT Testing & Evaluation SIG Newsletter, 5(1),* 13-16.

Brown, J. D. (2001c). Statistics Corner. Questions and answers about language testing statistics: What is two-stage testing? Shiken: *JALT Testing & Evaluation SIG Newsletter, 5*(2), 13-16.

Brown, J. D. (2001d). Statistics Corner. Questions and answers about language testing statistics: What is a point-biserial correlation coefficient? *Shiken: JALT Testing & Evaluation SIG Newsletter, 5(3),* 12-15.

Brown, J. D. (2002a). Statistics Corner. Questions and answers about language testing statistics: The Cronbach alpha reliability estimate. *Shiken: JALT Testing & Evaluation SIG Newsletter, 6*(1), 14-16.

Brown, J. D. (2002b). Statistics Corner. Questions and answers about language testing statistics: Extraneous variables and the washback effect. *Shiken: JALT Testing & Evaluation SIG Newsletter, 6(2),* 12-15.

Brown, J. D. (2002c). Statistics Corner. Questions and answers about language testing statistics: Distractor efficiency analysis on a spreadsheet. *Shiken: JALT Testing & Evaluation SIG Newsletter, 6(3),* 20-23.

Brown, J. D. (2003a). Statistics Corner. Questions and answers about language testing statistics: The coefficient of determination. *Shiken: JALT Testing & Evaluation SIG Newsletter, 7(1),* 14-16.

Brown, J. D. (2003b). Statistics Corner. Questions and answers about language testing statistics: Norm-referenced item analysis (item facility and item discrimination). *Shiken: JALT Testing & Evaluation SIG Newsletter, 7*(2), 16-19.

Brown, J. D. (2003c). Statistics Corner. Questions and answers about language testing statistics: Criterion-referenced item analysis (The difference index vs. the B-index). *Shiken: JALT Testing & Evaluation SIG Newsletter, 7(3),* 13-17

Brown, J. D. (2004a). Statistics Corner. Questions and answers about language testing statistics: Yates Correction. *Shiken: JALT Testing & Evaluation SIG Newsletter, 8(1),* 19-22.

Brown, J. D. (2004b). Statistics Corner. Questions and answers about language testing statistics: Test-taker motivations. *Shiken: JALT Testing & Evaluation SIG Newsletter, 8(2),* 16-20.

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

Brown, J. D. (2005b). Statistics Corner. Questions and answers about language testing statistics: Generalizability and decision studies. *Shiken: JALT Testing & Evaluation SIG Newsletter, 9*(1), 12-16.

Brown, J. D. (2005c). Statistics Corner. Questions and answers about language testing statistics: Characteristics of sound qualitative research. *Shiken: JALT Testing & Evaluation SIG Newsletter, 9*(2), 31-33.

Brown, J. D. (2006a). Statistics Corner. Questions and answers about language testing statistics: Resources available in language testing. *Shiken: JALT Testing & Evaluation SIG Newsletter, 10*(1), 21-26.

Brown, J. D. (2006b). Statistics Corner. Questions and answers about language testing statistics: Generalizability from second language research samples. *Shiken: JALT Testing & Evaluation SIG Newsletter, 10*(2), 24-27.

Brown, J. D. (2007a). Statistics Corner. Questions and answers about language testing statistics: Sample size and power. *Shiken: JALT Testing & Evaluation SIG Newsletter, 11*(1), 31-35.

Brown, J. D. (2007b). Statistics Corner. Questions and answers about language testing statistics: Sample size and statistical precision. *Shiken: JALT Testing & Evaluation SIG Newsletter, 11*(2), 21-24.

Brown, J. D. (2008a). Statistics Corner. Questions and answers about language testing statistics: The Bonferroni adjustment. *Shiken: JALT Testing & Evaluation SIG Newsletter, 12(1),* 23-28.

Brown, J. D. (2008b). Statistics Corner. Questions and answers about language testing statistics: Effect size and eta squared. Shiken: JALT Testing & Evaluation SIG Newsletter, *12*(2), 36-41.

Brown, J. D. (2009a). Statistics Corner. Questions and answers about language testing statistics: Principal components analysis and exploratory factor analysis—Definitions, differences, and choices. *Shiken: JALT Testing & Evaluation SIG Newsletter, 13(1)*, 26-30.

Brown, J. D. (2009b). Statistics Corner. Questions and answers about language testing statistics: Choosing the right number of components or factors in PCA and EFA. *Shiken: JALT Testing & Evaluation SIG Newsletter, 13*(2), 19-23.

Brown, J. D. (2009c). Statistics Corner. Questions and answers about language testing statistics: Choosing the right type of rotation in PCA and EFA. *Shiken: JALT Testing & Evaluation SIG Newsletter, 13(3),* 20-25.

Brown, J. D. (2010a). Statistics Corner. Questions and answers about language testing statistics: How are PCA and EFA used in language research? *Shiken: JALT Testing & Evaluation SIG Newsletter*, *14*(1), 19-23.

Brown, J. D. (2010b). Statistics Corner. Questions and answers about language testing statistics: How are PCA and EFA used in language test and questionnaire development? *Shiken: JALT Testing & Evaluation SIG Newsletter, 14*(2), 22-27.

Brown, J. D. (2011a). Statistics Corner. Questions and answers about language testing statistics: Likert items and scales of measurement. *Shiken: JALT Testing & Evaluation SIG Newsletter, 15*(1), 10-14.

Brown, J. D. (2011b). Statistics Corner. Questions and answers about language testing statistics: Confidence intervals, limits, and levels? *Shiken: JALT Testing & Evaluation SIG Newsletter, 15*(2), 23-27.

Brown, J. D. (2012a). Statistics Corner. Questions and answers about language testing statistics: What do distributions, assumptions, significance vs. meaningfulness, multiple statistical tests, causality, and null results have in common? *Shiken Research Bulletin, 16*(1), 28-33. Brown, J. D. (2012b). Statistics Corner. Questions and answers about language testing statistics: How do we calculate rater/coder agreement and Cohen's kappa? *Shiken Research Bulletin,* 16(2), 30-36.

Brown, J. D. (2013a). Statistics Corner. Questions and answers about language testing statistics: Chi-square and related statistics for 2 x 2 contingency tables. *Shiken Research Bulletin, 17*(1), 33-40.

Brown, J. D. (2013b). Statistics Corner. Questions and answers about language testing statistics: Solutions to problems teachers have with classroom testing. Shiken Research Bulletin, 17(2), 27-33.

Brown, J. D. (2014a). Statistics Corner. Questions and answers about language testing statistics: Differences in how norm- referenced and criterion-referenced tests are developed and validated. *Shiken Research Bulletin, 18*(1), 29-33.

Brown, J. D. (2015a). Statistics Corner. Questions and answers about language testing statistics: Testing intercultural pragmatics ability. *Shiken Research Bulletin, 19(1),* 42-47.

Brown, J. D. (2015b). Statistics Corner. Questions and answers about language testing statistics: Characteristics of sound quantitative research. *Shiken Research Bulletin, 19*(2), 24-28.

Brown, J. D. (2016a). *Statistics corner: Questions and answers about testing statistics*. JALT.

Brown, J.D. (2016b). Background to this book. In J.D Brown, *Statistics corner: Questions and answers about testing statistics*. (pp. xii-xvi). JALT.

Brown, J. D. (2016c). Statistics Corner. Questions and answers about language testing statistics: Characteristics of sound mixed methods research. *Shiken Research Bulletin, 20*(1), 21-24.

Brown, J. D. (2016d). Statistics Corner. Questions and answers about language testing statistics: Consistency of measurement categories and subcategories. *Shiken Research Bulletin, 20(2),* 50-53.

Brown, J. D. (2017a). Statistics Corner. Questions and answers about language testing statistics: Consistency in research design: Categories and subcategories. *Shiken Research Bulletin, 21*(1), 23-28.

Brown, J. D. (2017b). Statistics Corner. Questions and answers about language testing statistics: Developing and using rubrics: Analytic or holistic? *Shiken Research Bulletin, 21*(2), 20-26.

Brown, J. D. (2018a). Statistics Corner. Questions and answers about language testing statistics: Developing rubrics: What steps are needed? *Shiken Research Bulletin, 22*(1), 7-13.

Brown, J. D. (2018b). Statistics Corner. Questions and answers about language testing statistics: Calculating reliability of dictation tests: Does K-R21 work? *Shiken Research Bulletin, 22*(2), 14-19.

Brown, J. D. (2019a). Statistics Corner. Questions and answers about language testing statistics: What is assessment feedback and where can I find out more about it? *Shiken Research Bulletin, 23*(1), 45-49.

Brown, J. D. (2019b). Statistics Corner. Questions and answers about language testing statistics: Overall English proficiency (whatever that is). *Shiken Research Bulletin, 23*(2), 43-47.

Brown, J. D. & Yamashita, S. O. (1995). English language entrance examinations at Japanese universities: What do we know about them? *JALT Journal*, 17(1), 7-30.

Fisher, R. A. (1971). The design of experiments (8th ed.). Hafner. Reproduced in J. H. Bennett (Ed.) (1995) *Statistical methods, experimental design, and scientific inference*. Oxford University.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 33-45). Lawrence Erlbaum Associates.