SRB

SHIKEN RESEARCH BULLETIN

Volume 16 • Number 1 • May 2012

Contents

Testing and Evaluation SIG

Foreword

Aaron Olaf Batty
TEval SIG Publications Chair,
Shiken Research Bulletin *General Editor*

A fond farewell

As you have no doubt already noticed, this issue of *Shiken* marks some big changes to the publication. Our former General Editor and TEval Publications Chair, Tim Newfields, has passed the reigns of the publication to me and a small team of new faces after many, many years of dedicated, solitary service. For most readers, the only *Shiken* you have known has been the results of Tim's extraordinary service to the SIG, and it is with great respect and humility that we, the new editorial board, receive the duties of producing this publication. We hope to continue in Tim's tradition and ensure that this publication remains the most interesting and useful testing publication for language testers in Japan.

Introductions

The new General Editor is myself, Aaron Olaf Batty, of Keio University, Shōnan-Fujisawa Campus. I am joined by Jeffrey Stewart, of Kyūshū Sangyō University as Associate Editor, and by Aaron Gibson and Jeff Durand as Assistant Editors. The oddity of having four members with only two first names is not lost on us.

A new name

The first and most obvious change is that of the name. This publication has always striven to primarily feature research and research-oriented pieces of direct relevance and interest to the Japanese language testing community. For that reason, the editorial board, with the blessing of the SIG officers, has renamed the publication *Shiken Research Bulletin,* a name that accentuates this editorial position, and elevates the publication above a "newsletter." With this name comes a new abbreviation (*SRB*), and a new logo.

A new look

SRB articles now follow a standardized template that seeks to enhance readability and ease production. In addition, printing of the newsletter has been outsourced to a professional printer. We hope you like it.

More to come…

We have many more changes planned and in progress, especially with regards to the web version of *SRB,* and with the SIG website. Stay tuned.

We hope you enjoy this issue of *SRB,* and from myself and the new Editorial Board:

どうぞ宜しくお願い致します。

A Bayesian alternative to null hypothesis significance testing

John Eidswick
johneidswick@hotmail.com
Konan University

Abstract

Researchers in second language (L2) learning typically regard "statistical significance" as a benchmark of success for an experiment. However, because this statistic indicates nothing more than the probability of data sets occurring given the essentially impossible condition that the null hypothesis is true, it confers little of practical or theoretical importance. Significance is also the source of widespread misinterpretation, including confusion of significance with effect size. Critics of NHST assert that alternative approaches based on Bayes' theorem are more appropriate for hypothesis testing. This paper provides a non-technical introduction to essential concepts underlying Bayesian statistical inference, including prior probabilities and Bayes factors. Common criticisms of NHST are outlined and possible benefits of Bayesian approaches over NHST are discussed.

Introduction

In this article, I provide an overview of Bayesian statistics and contrast it with null hypothesis significance testing (NHST). I also describe criticisms often expressed about NHST (e.g. in Cohen, 1994) and reasons that Bayesian statistics might be a suitable alternative for analyses in second language (L2) learning research. I will also outline concepts important to Bayesian approaches, such as prior probability distributions and Bayes factors.

Perhaps the best way to introduce Bayesian statistics is by way of an example. Research has demonstrated that the motivational variable *interest* has a powerful effect on processes important to reading comprehension (see Hidi & Renninger, 2006, for a review). A researcher wants to learn whether interest influences comprehension in L2 reading in a comparable way as occurs in first language (L1) contexts, so she has a group of 25 students read an interesting and a boring story and take comprehension tests. She checks for differences between the test scores by using a *t* test.

Researchers use *t* tests to compare two groups of data produced under different conditions to determine the probability that no difference exists between them beyond random variation. The hypothesis that no difference exists is called the null hypothesis (H_0). Data from a *t* test consists of an independent variable (IV) that is manipulated and a dependent variable (DV) that might be affected by the IV.

The probability (the *p* value) that a *t* statistic of the size produced by the test would occur given that H_0 is correct is calculated. A *p* value of less than .05 would mean that if we were to repeat this test 100 times, a statistic of this size or higher would result by random chance fewer than five times (see Figure 1). In this case the results are considered "significant" and the researcher rejects the null hypothesis.

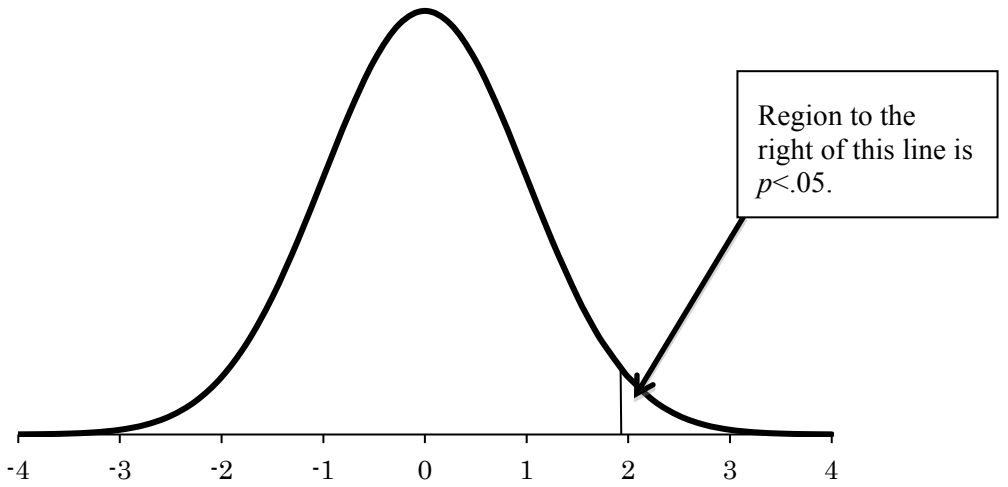

***Figure 1*. Idealized distribution of scores with threshold *p*=.05 marked. If the *t* statistic falls to the right of this line, the mean difference between H_0 and H_1 is considered statistically significant.**

Note that space constraints do not permit detailed discussion in this article of one- and two-tailed tests, *t* distributions, degrees of freedom, confidence intervals, effect sizes, or statistical power, but these are also important aspects of NHST. Readers unfamiliar with these concepts are encouraged to read explanations that can be found in many introductory statistics textbooks (e.g. Field, 2009).

Our researcher performs a *t* test on the data (in reality, the data was produced using a random number generator for normal distributions at Wessa, 2008). Table 1 shows the descriptive statistics and the results are provided in Table 2.

Table 1. Descriptive Statistics for Boring and Interesting Text Conditions (*n* = 25)

Text Group	N	M	SD	SE
Boring	25	10.24	2.13	.43
Interesting	25	10.92	3.76	.75

Table 2. *T* Test Results for Boring and Interesting Text Conditions

Df	SED	MD	T	p (one-tailed)
24	.489	.68	1.69	.098

Note: *p*<.05.

As can be seen, the results are insignificant at *p*<.05, because .098 is larger than .05. The researcher therefore is inclined not to reject the null hypothesis. A colleague of our researcher, on a lark, does the same experiment with a very similar group of students and adds the scores to the original data. The *t* test is performed again, now with an *n* size of 50. The new descriptive statistics and *t* test results are provided in Tables 3 and 4.

Table 3. Descriptive Statistics for Boring and Interesting Text Conditions (*n* = 50)

Text Group	N	M	SD	SE
Boring	50	10.24	2.10	.30
Interesting	50	10.92	3.72	.53

Table 4. T Test Results for Boring and Interesting Text Conditions (*n* = 50)

Df	SED	MD	T	p (one-tailed)
49	.342	.68	1.99	.026*

Note. *$p<.05$.

As we can see, the descriptives have changed little, but the results of the *t* test are now significant. By convention, these results are now considered publishable, despite the fact that for all practical purposes they are identical to those of the previous experiment. This poses a serious dilemma for our researchers. Should they reject or not reject the null hypothesis? In order to get published, should they favor the second results and pretend those of the original study did not occur? The source of the dilemma lies in the fact that finding significance is reliant on statistical power, which is related to sample size. Such is the relationship between *N* size and significance that in the case of a large enough number of cases, finding significance is all but certain, irrespective of an actual experimental effect. This is one of several problems described by critics of NHST.

Criticism of null hypothesis significance testing

Criticism of NHST can be divided into two categories: 1) weaknesses of NHST as an evaluative tool, and 2) misinterpretations of what NHST results mean.

Weaknesses

Problematic qualities of NHST cited by critics include:

- The primacy of significance. Editors of L2 learning academic journals tend to view significance as the mark of success, but using the *p* value as a Litmus test for "success" is problematic. As noted, variations of sample size can change the likelihood of finding significance, a characteristic that leads some statisticians to argue that NHST *p* values do not qualify as measures of statistical evidence, much less success, because identical *p* values do not convey identical levels of evidence when sample sizes differ (Wagenmakers, 2007).
- The primacy of significance thresholds. R.A. Fisher, whose work underpinned the development of the *p* value, did not himself regard *p* values as rigid cut-off points (Salsburg, 2001). In fact, Fisher's selection of these benchmarks was at least somewhat arbitrary; he identified *p* value thresholds in his book *Statistical Methods for Research Workers* (1925) by providing critical values tables, which were limited to .05, .02, and .01, to "save space" (Field, 2009, p. 51). As Abelson (1997) commented, "Literal insistence on the .05 level is as silly as would be other arbitrarily rigid quality standards for research results, like 30% generality, or more interestingness than three quarters of the existing literature" (p. 14).
- Power issues. Statistical power refers to the probability that a test can detect an effect. Adjustments made in analyses to reduce the chances of incorrectly rejecting H_0 (Type I

errors) or incorrectly failing to reject H_0 (Type II errors), inevitable in a range of NHST-based tests, involve a loss of statistical power.

Misconceptions

Misconceptions persist regarding what NHST in fact measures and what its results mean. Some are outlined below.

- The meaning of significance. It is tempting to conclude that statistical significance indicates that the null hypothesis is false and the alternate hypothesis is true, but this is not correct. The finding of significance really means nothing more than the researcher is inclined to reject the null hypothesis based on a low probability (defined by a somewhat arbitrarily chosen threshold) that the data at hand would occur randomly in many recursions of it. Significance does not mean the null hypothesis is formally invalid. Rejecting H_0 does not prove H_1, but instead merely offers an indirect and rather flimsy indicator of support for it.
- The meaning of lack of significance. Because NHST purports to test the hypothesis that the null hypothesis is true, one might well be inclined to infer that an "insignificant" finding means just that: the null hypothesis is true. In fact, situations where mean differences are literally zero virtually never occur in the real world.
- Substitution of a conditional probability for its inverse. The notion that the probability of certain data given that H_0 is true, that is, $P(D|H_0)$, is equivalent to the probability that H_0 is true given certain data, or $P(H_0|D)$, is intuitively appealing, but the difference between these two becomes clear when contemplating the probability of having a runny nose given the condition of having the flu versus having the flu given the condition of having a runny nose; one can have a runny nose for many reasons besides the flu.
- The idea that "significance" means "importance". Significance does not refer to the magnitude of an experimental effect. Reporting effect sizes in published studies would help clarify this misunderstanding, but few L2 researchers do so.

While the issues outlined above might create obstacles to rigorous hypothesis construction and testing, critics of NHST assert that a solution to many of these problems lies in Bayesian statistics.

Bayesian statistics

Bayesian statistical approaches are drawn from the work of 18th century mathematician Thomas Bayes. Bayesian and classical statistical approaches differ crucially in two areas. The first relates to how analyses are interpreted. Frequentist approaches like NHST produce *p* values that estimate the likelihood that the data would occur given that the null hypothesis is true. Rather than stating a "cut point" after which one hypothesis is chosen over another, Bayesian analyses result in probability values that are used to compare the relative support for one hypothesis over another. In short, frequentists seek significance and Bayesians seek probability support for a hypothesis. The second area where the two kinds of approaches differ is the formal use of prior information. Frequentists ignore what was previously known about the experimental condition when conducting a new experiment, but incorporating this prior information into future analyses is an essential part of Bayesian approaches. Prior information is incorporated by using the feature which most distinguishes Bayesian statistics, the *prior probability*.

Thomas Bayes's solution to a problem of "inverse probability" (e.g., estimating the unknown likelihood of an event happening given the known likelihood of a certain condition) contained a description of the formula which has come to be known as Bayes' theorem. As noted, a distinctive feature of the formula is the prior probability (the *prior*). In a basic application of Bayes' formula, the multiplied product of the prior and the probability of data given a certain parameter is divided by the probability of the parameter defining the sample space to produce a *posterior probability* (the *posterior*):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where A is the parameter under investigation and B is the data.

Let us make our introduction to Bayes less abstract with a simple example (adapted from Bonilla, 2011; for another non-technical example of Bayes' formula, see Yudkowsky, 2003). The data used in the first *t* test could be categorized as individuals who passed and failed the comprehension test. If we consider test scores of 12 (60%) or higher as "pass" and those below as "fail," then 6 of 25 students (24%) passed the test in the boring text condition, and 15 of 25 students (60%) passed in the interesting text condition. To understand how Bayes' formula works, it is useful to concentrate on how the *Interest* condition relates to the *Pass* scores (see Figure 2).

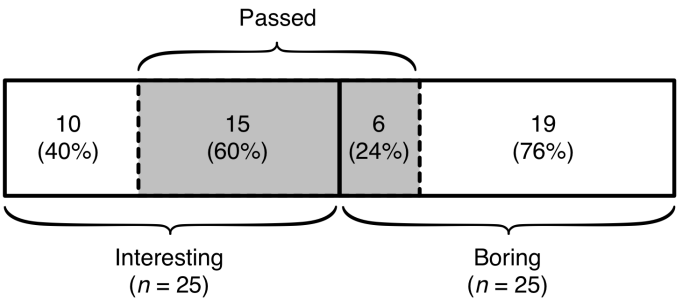

Figure 2. **Interesting and boring text conditions divided into passed (shaded areas) and failed (unshaded areas) results on comprehension tests.**

The rectangle on the left indicates the Interesting condition, the rectangle on the right the boring condition, The shaded areas indicate the proportions of students in each condition that passed the test. The shaded rectangle to the left of the center dividing line represents the intersect of students who found the text interesting *and* passed the test; it is denoted as *Int* ∩ *Pass,* which can be read as "*Int* and *Pass* happen together". Likewise, the shaded portion of the "Boring" condition indicates the intersect of students who found the text boring with those who passed the test (*Bor* ∩ *Pass*).

What is the probability that a student passed the test if he read the interesting text? In conditional probability notation, this is signified by *P*(*Pass* | *Int*), which is read, "the probability of the event of a student's passing given the event that the student read the interesting text". We can think of this as the answer to the question, "how much of the *Interesting* rectangle is accounted for by the shaded *Passed* area?" The answer is already given as 60%.

This is not what the researcher really wants to know, however. The researcher is interested not in the probability of passing given that the text is interesting, but in the probability of the text being interesting given that the student passed. This is represented by *P*(*Int* | *Pass*), which means, "the

probability the text was interesting given that the test was passed". We can easily calculate this information from Bayes' formula as follows:

$$P(Int|Pass) = \frac{P(Pass|Int)P(Int)}{P(Pass)}$$

The values in the numerator are already known—the probability a student passed given that he thought the text was interesting is 0.60:

$P(Pass|Int) = .60$

and the probability that the student found the text interesting is 0.50, as half the scores came from the interesting condition:

$P(Int) = .50$

The shaded *Pass* area has two parts: (A) the section contributed by the *Int* condition, and (B) the section contributed by the non-*Int* (boring) condition. To calculate $P(Pass)$, these two parts are added.

In calculating (A), we are asking, "how much of the *Pass* area is made up of passing scores in the context of the *Interest* rectangle? The probability of a passing score within the *Interest* rectangle, *P*(*Pass* | *Int*), is .60. To find out how much of this probability contributes to the *Pass* area, we simply multiply it by *P*(*Int*), which is the probability of the student finding the text interesting, or 50%:

.60 × .50 = .30.

Likewise, in calculating (B), we are asking "how much of the *Pass* area is comprised of passing scores in the context of the *Boring* condition?" Mathematically, this is *P*(*Pass* | *B*) × *P*(*B*). We know that 24% of the students who read the boring text passed the test and that 50% of the scores came from this condition. Therefore, (B) is calculated:

.24 × .50 = .12.

The denominator in Bayes formula, $P(Pass)$, then, is:

.30 + .12 = .42.

Plugging our values into Bayes' formula, we obtain:

$$P(Int|Pass) = \frac{.30}{.42} = .71$$

Therefore, if we randomly draw a student with a passing test score from this group, there is a 71% chance that he read the interesting text. If we were to calculate *P*(*B* | *Pass*) using Bayes formula, we would obtain the remaining percentage of this region, that is, 29%. Of course, since probability of a given space must add up to 1, we could also simply subtract .71 from 1 to derive .29. We can now add these probabilities to the previous diagram:

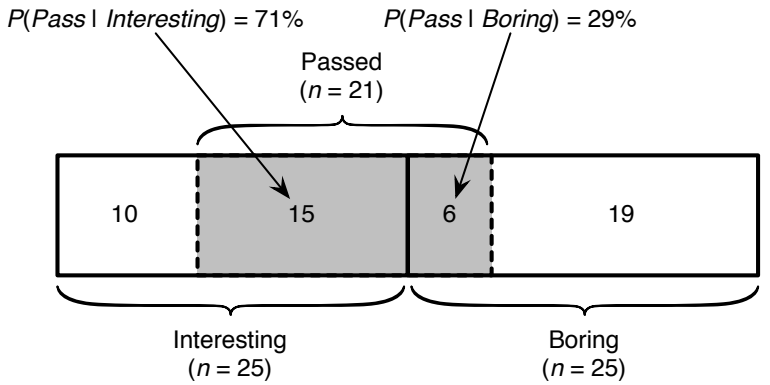

Figure 3. **Bayes formula provides probabilities of passing conditions.**

Although in this simple example the difference between pass and fail percentages is stark; however, it should be noted that the results are not always so obvious. If far more students found the texts boring than interesting, it is possible that a student that passed will *still* have a higher likelihood of having thought his book was boring, even if a much higher proportion of students who thought the texts were interesting passed. Bayes' rule adjusts for differences in sample sizes between conditions when calculating these probabilities.

Prior probabilities

The simple example above might be useful for illustrating the basic dimensions of Bayes theorem, but it elides over some important points, chiefly to do with the assignment of the prior probability. Unlike classical statistics, Bayesian approaches enable researchers to include relevant prior information in formal experimentation. If previous research indicates one outcome is more likely than another, a Bayesian can integrate this information into his hypothesis formation and testing (Figure 3). The revised probability resulting from the new experiment can then influence the selection of priors used in subsequent investigations to further refine probability estimations in support of one hypothesis or other. The prior probability is a summary of a researcher's belief about the outcome of a given experiment.

In the example above, for simplicity, the prior, $P(I)$, was given as a known value and as a simple mean, but in a normal Bayesian analysis, the prior would be designated based on personal belief of the researcher. This personal belief could be drawn from previous research, or even just the researcher's conjecture. Since Bayesians, unlike frequentists, regard unknown values under investigation as random variables (that is, variables that manifest as values with certain probabilities), the conjectured outcomes for these values, expressed by the prior, take the form of probability distributions, indicated concretely by parameters like mean, standard deviation, and range. If prior information is lacking, the researcher can use a prior that expresses a high degree of uncertainty. High uncertainty can be related by designating a prior with a large standard deviation and by maximizing the range (Klugkist & Mulder, 2008). When probability distributions are used in Bayesian procedures, the calculations become much more complicated than those of our example. They are accomplished using calculus and sophisticated algorithms (such as Markov Chain Monte Carlo) that require a computer to generate. The good news is that software for using these procedures is available.

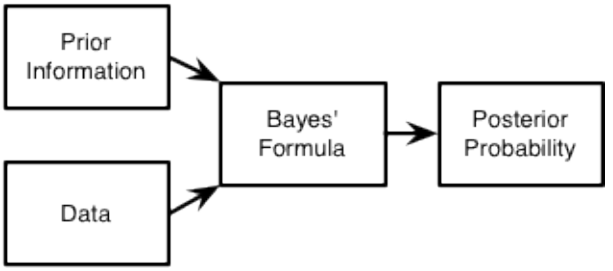

Figure 4. **Bayesian approaches combine prior information, in the form of a prior probability, to produce an updated view of phenomena, in the form of a posterior probability (adapted from Stevens, 2009).**

Why use a prior?

To understand why Bayesians incorporate prior information into statistical inference, it is helpful to understand the quite different perceptions frequentists and Bayesians have of probability and the different goals the two kinds of researchers have in hypothesis testing. For frequentists, probability is the likelihood that a certain unknown (and ultimately unknowable) value lies within a distribution of values drawn from many samplings of a population, with the goal of analyses being to estimate whether sampled data would occur less than five percent of the time ($p < 0.05$) given the null hypothesis is true. For Bayesians, probability is conceived as a degree of personal belief which can be refined by confrontation with real-world evidence. In Bayesian statistics, the goal is to modify a given state of knowledge about a phenomenon by connecting it to data; to do so without concretely representing the state of knowledge would be impossible. This existing state of knowledge is represented using the prior. Moreover, because the prior summarizes researcher belief about experimental outcomes, it can also be considered an expression of a hypothesis, a prediction subject to modification given new information collected during the new experiment.

To make this clearer, let's look at another example. Our first researcher, devastated by her *t* test debacle, throws away her data. A second researcher, a Bayesian, discovers her data while rooting through the trash bin. He decides to analyze it using a Bayesian approach.

Our Bayesian researcher considers some choices for a prior. In the absence of much prior information or a defined hypothesis, he might choose a prior that indicates only that a range exists in scores, from 0 to 20, with each score having equal probability of occurring. The X-axis of the uniform prior (Figure 4) shows the range of comprehension test scores from zero to 20, and the Y-axis shows the probability of those scores according to the prior. This kind of uniform, approximately objective, prior conveys much uncertainty and provides little information, so the data will dominate the calculation of the posterior probability.

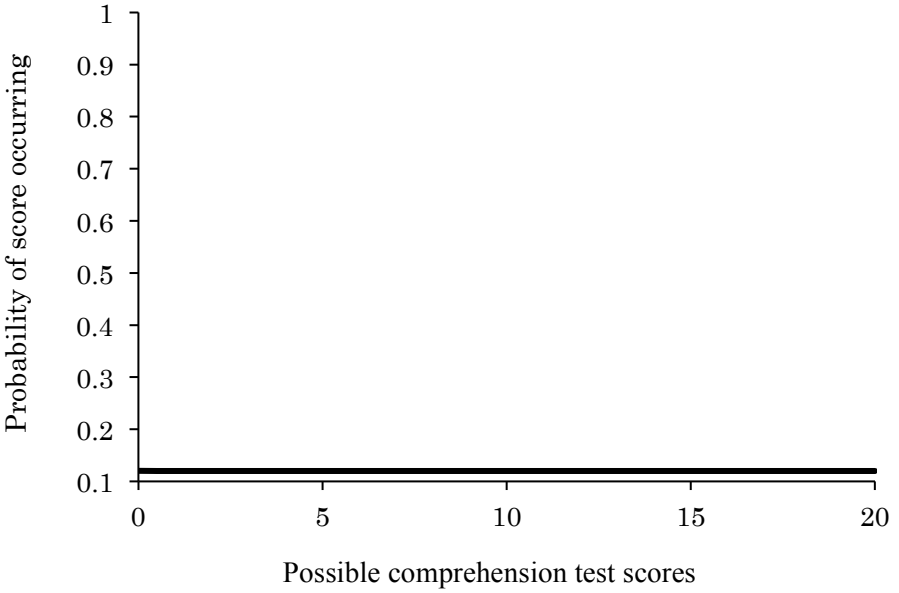

***Figure 4*. Uniform prior where each score on the comprehension test has an equal probability of occurring. While this prior would be approximately objective, it would also likely be an unrealistic representation of the data.**

Our researcher might view a uniform prior as unrealistic, since data are unlikely to be flatly distributed. Also, while the results of the scavenged *t* tests were contradictory in terms of significance, they do suggest a modest degree of variance attributable to the interesting text condition. Our researcher gleans further in a literature review that related (fictional) studies indicate that interest contributes about 5% to increases in text comprehension. He could assign a prior with a mean score 5% higher than would occur by chance (i.e. one point higher than a mean of 10) with a standard deviation of 5. The standard deviation for a normal distribution can be estimated by dividing the highest extreme of the range of scores, in this case 20, by four. Figure 5 shows the distribution of this subjective prior.

The X-axis shows the range of comprehension scores, and the Y-axis shows the probability of the scores occurring. For example, a score of 11 (the mean) would have a probability of approximately .16 or 16% of occurring whereas a score of 5 would have about .01 or 1% chance of occurring. Using this subjective prior would involve the meeting of the hypothesized outcome represented by the prior (interest influences comprehension positively by a predicted amount, with a predicted degree of dispersion) with the data. The probabilities predicted by this prior would be somewhat higher than those predicted to occur by chance, so while this prior is subjective, it is also quite conservative.

Our Bayesian decides to use the more informative prior to test his hypothesis. To aid in the calculations, and to compare the relative support of his hypothesis with that of the null, he uses a *Bayes factor*.

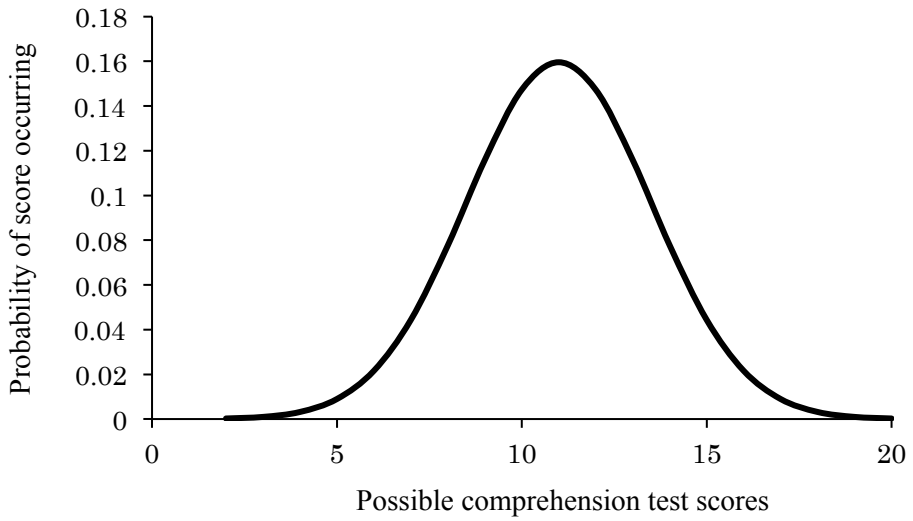

Figure 5. **Prior with a normal distribution and a mean of 11. This somewhat subjective prior would reflect a hypothesized distribution based on previous research.**

Bayes factors

In a Bayesian version of a *t* test, the probability of H_0 and its alternative are compared to produce a statistic called a Bayes factor (BF). Put simply, the BF is a ratio that compares the likelihood of one model over another, thereby showing the relative support for the researcher's hypothesis versus another hypothesis (which may or may not be the null).

Interpretation of the BF is straightforward. For example, a BF of 4 for H_1 versus H_0 indicates support for H_1 is 4 times that of H_0. A BF of .5 provides two times the support for H_0 than for H_1 (Klugkist, 2008). Bayes factors between .3 and 3 do not provide much evidence to differentiate the two hypotheses (Jeffreys, 1939, 1998).

In order to run the necessary calculations, the researcher uses an online Bayes factor calculator, provided at the following link (Dienes, 2008):

http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/bayes_factor.swf

The calculator provides a limited range of priors templates for calculating a simple BF for a Bayesian "*t* test" that shows the relative support for the null hypothesis, as manifested by a population value with a mean of 0, and a hypothesis, as expressed by the mean differences and distributions assigned by the researcher (For supporting explanation, see Dienes, 2011).

Some simple modifications to our researcher's data are required to use the calculator. Our researcher chooses a normal distribution option and enters a mean difference of 5% with a range of 1% to 10% and a standard deviation of 2.5. With the *t* test data ($n = 50$), the BF produced is .45. This indicates slightly more support for the null versus the interest hypothesis. This contrasts with the finding of statistical significance in the second *t* test, and that our researcher's interpretation of the BF involves neither rejecting nor failing to reject the null, but instead making inferences based on the comparative likelihoods of H_0 and H_1.

Given this result, the researcher would be inclined to align his confidence somewhat away from the findings of other research in L1-based contexts. He might give careful consideration to the special characteristics of L2 readers and design future studies to explore in a more nuanced way potential effects of interest on reading comprehension. In the service of using accumulated experiences to update and refine knowledge, the results of a Bayesian analysis are used to contribute to new hypotheses and to shape the priors assigned in future studies. Note that this kind of statistical reasoning is essentially different from that of many researchers interpreting the results of a NHST. In the case of the example *t* tests, our first researcher either rejects the null hypothesis (with the higher *n* size) or does not reject the null hypothesis (with the lower *n* size), and then draws only peripherally related conclusions.

Bayesian benefits

Several qualities of Bayesian statistics might render them useful in L2 learning research, including:

- Flexibility. Bayesian approaches permit direct comparisons between multiple hypotheses by incorporating inequality constraints; null hypothesis-based comparisons of multiple hypotheses require secondary procedures, such as post hoc tests, which can result in reduced statistical power and can yield mutually inconsistent results.
- Protection against fallacious "significance". Power and *N* sizes are not irrelevant to Bayesian approaches, but, unlike NHST, high numbers of cases do not inevitably result in something akin to "significance." Instead, in a Bayesian *t* test where the null is approximately correct, higher *N* sizes drive the BF toward zero (Dienes, 2011).
- Validity: Bayesian methods directly address questions researchers are trying to answer. Unlike frequentists, who test "'nothing is going on' versus 'something is going on but I don't know what'" (Boelen & Hoijtink, 2008, p. 10), Bayesians ask, "what is the chance my hypothesis is true given the evidence?"
- Possible: A variety of Bayesian software packages are available, some reasonably user-friendly. For example, the Bayesian Inequality and Equality Model Selection (BIEMS) program (Mulder, Hoijtink, & de Leeuw, 2012; Mulder, Hoijtink, & Klugkist, 2010; Mulder, Klugkist, van de Schoot, Meeus, Selfhout, & Hoijtink, 2009) is available for free and has a Windows user interface. For a thoroughgoing description of available software packages for Bayesian approaches, see Hoijtink, (2012).
- Objective: Perhaps the most pervasive criticism of Bayesian approaches relates to the prior, which entails, it is believed, a subjective and therefore biased decision by the researcher. However, vague or uninformative priors can be assigned which are approximately objective.

Conclusion

Researchers in L2 learning use NHST almost exclusively. However, many researchers are unfamiliar with the limitations of NHST and unaware that alternative procedures, such as those related to Bayes' theorem, exist. Despite the growing wealth of explanatory materials and availability of software by which even non-statisticians can avail themselves to Bayesian statistical methods, to date, no researcher to my knowledge has attempted to use these potentially advantageous procedures in research focused on L2 learning.

References

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8,* 12-15.

Boelen, P. A., & Hoijtink, H. (2008). Illustrative psychological data and hypotheses for Bayesian inequality constrained analysis of variance. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp.7-26). New York: Springer.

Bonilla, O. (2011). Visualizing Bayes' theorem. *Math115.com*. Retrieved from http://math115.com/2011/02/visualizing-bayes%e2%80%99-theorem/

Cohen J. (1994) The earth is round ($p < 0.05$). *American Psychologist, 49*(12), 997–1003.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference* [Supplemental material]. Hampshire, England: Palgrave Macmillan. Retrieved from http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/

Dienes, Z. (2011). Bayesian Versus Orthodox Statistics: Which Side Are You On? *Perspectives on Psychological Science 6*(3) 274–290 DOI: 10.1177/1745691611406920

Field, A. (2009). *Discovering statistics using SPSS*. London: Sage.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Hidi, S. & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*, 111-127.

Hoijtink, H. (2012). Informative hypotheses: Theory and practice for behavioral and social scientists. London: CRC Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795.

Klugkist, I. & Mulder, J. (2008). Bayesian estimation for inequality constrained analysis of variance. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 27-52). New York: Springer.

Klugkist, I. (2008). Encompassing prior based model selection for inequality constrained analysis of variance. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp.53-83). New York: Springer.

Jeffreys, H. (1939, 1998). *Theory of probability*. 3rd Ed. Oxford: Oxford University Press.

Mulder, J., Hoijtink, H. & de Leeuw, C. (2012). BIEMS: A Fortran90 program for calculating Bayes factors for inequality and equality constrained models. *Journal of Statistical Software, 46*(2), 1-39.

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference, 140*, 887-906.

Mulder, J. Klugkist, I., van de Schoot, R., Meeus, W., Selfhout, M., and Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology, 53*, 530-546.

Stevens, J. W. (2009). What is Bayesian statistics? Retrieved from http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/What_is_Bay_stats.pdf

Wagenmakers, E. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779-804.

Wessa, P. (2008). Random number generator for the normal distribution (v1.0.8). Free Statistics Software (v1.1.23-r7), Office for Research Development and Education. Retrieved from http://www.wessa.net/rwasp_rngnorm.wasp/

Yudkowsky, E. S. (2003). *An intuitive explanation of Bayes' theorem*. Retrieved from http://yudkowsky.net/rational/bayes

Does IRT provide more sensitive measures of latent traits in statistical tests? An empirical examination

Jeffrey Stewart
jeffjrstewart@gmail.com
Kyushu Sangyo University, Swansea University

Abstract

It has been frequently stated that Rasch and Item Response Theory produce interval-scale measures where raw scores can only provide ordinal measures, and that therefore, researchers should choose Rasch/IRT measures when selecting variables for common statistical tests (Wright, 1992; Harwell & Gattie, 2001). In this study, this claim is empirically examined by conducting Pearson Correlations and ANOVAs on two data sets using raw scores, Rasch Person Measures and 2-Parameter IRT ability estimates, in order to determine if results differed as a consequence. Raw Scores and Rasch Person Measures were very highly correlated, and lead to extremely similar results in all cases. For a well-constructed, reliable test the same was true of 2PL ability estimates. However, in cases where the test has middling to poor reliability, 2PL ability estimates appear to produce a somewhat more sensitive measure of a latent trait than raw scores, which can result in meaningful differences in statistical tests.

Introduction

Noteworthy proponents of Rasch Measurement and Item Response Theory (IRT) have argued that raw test scores used under Classical Test Theory (CTT) are ordinal and non-linear in nature, and therefore not suitable for use in "normal" (i.e., parametric) statistics (e.g. Wright, 1992). The theoretical argument underlying this claim has led proponents of Rasch Measurement and Item Response Theory to warn against the use of raw scores from psychological tests as variables in experiments and statistical analyses. As Harwell and Gattie (2001) wrote in Review of Educational Measurement, "educational researchers frequently employ ordinal-scaled dependent variables in statistical procedures that assume that these variables possess an interval scale of measurement . . . . data possessing an ordinal scale will not satisfy the assumption of normality needed in many statistical procedures and may produce biased statistical results that threaten the validity of inference" (p. 105). The conversion of test data to raw scores was illustrated using the Rasch model.

Harwell and Gattie explain how to rescale the theoretically ordinal data produced by raw scores into theoretically interval data using item response theory, but they do not illustrate the usefulness of the rescaled data in a common experimental design, for example, comparing the use of Rasch measures in a t-test or ANOVA rather than raw scores. When using scores on a psychometric test as a dependent variable, does the substitution of raw scores for Rasch/IRT Measures have a practical effect on a statistical analysis?

Although theoretical arguments for the superiority of Rasch and Item Response Theory over Classical Test Theory (CTT) abound, there has been a relative paucity of empirical research regarding the practical benefits of using ability estimates calculated under item response theory over raw scores in this regard. Noting this, Fan (1998) argued, "Theoretical models are important

in guiding our research and practice. But the merits of a theoretical model should ultimately be validated through rigorous empirical scrutiny" (p. 15). In Fan's study comparing CTT, Rasch, 2PL IRT and 3PL IRT estimates of item difficulty and person ability, he found that all measures were highly comparable and very highly correlated. Though he acknowledged the usefulness and practical advantages of Rasch and IRT for applications such as item banking and computer adaptive testing, he concluded that differences between CTT and forms of IRT were overstated, and that despite theoretical arguments to the contrary, the actual differences between the two approaches were minimal. Various subsequent studies (e.g., MacDonald & Paunonen, 2002; Progar, Socan & Pec, 2008) have essentially confirmed these findings.

Of course, high Pearson correlations do not necessarily indicate that IRT estimates and raw scores are identical in every respect. As Linacre (1998) explains, even if two measures are perfectly correlated, the length of intervals between scores can vary greatly. Since these intervals are accounted for by logits, change scores of the same numerical value can be considered equivalent, though the same often cannot be said of raw scores (Embretson & Reise, 2000). Still, the strong similarities are discouraging, and do little to encourage researchers outside of psychometrics to move from raw scores to modern test theory approaches.

Suppose a researcher in SLA wishes to compare the effects of two teaching approaches on test scores. Theoretical arguments aside, and even supposing a given IRT model is shown to have better fit to the data using a statistic such as the BIC, ultimately do the approaches in scoring methods genuinely differ enough in practice that one should be preferred over others? The question is of central importance to any language acquisition researcher who has considered using Rasch or IRT measures for tests or surveys, because a common assumption is that the use of such models will improve the measurement properties of the research instruments they are applied to, and that by extension the likelihood of detecting a statistically significant difference between treatments in a common statistical analysis will increase. If this is not the case, it is more difficult for researchers outside of psychometrics and educational assessment to justify the time and expense of adapting to modern test theory and buying and learning to operate the software programs required to operationalize it.

Research Questions

Consequently, this short paper will attempt to answer three questions:

If used in place of raw scores, do Rasch/2PL IRT ability estimates improve the correlation of one language test to another?

Are the results of an ANOVA noticeably different if Rasch/2PL IRT ability estimates are used as the dependent variable rather than raw scores?

Do results differ if the above experiments are conducted using less reliable tests?

Method and Analyses

Data from two tests were used in this study, one with a high reliability (a Cronbach Alpha of 0.91) and one with a relatively low reliability (a Cronbach Alpha of 0.75). This was done to test if Rasch and IRT ability estimates were effective in reducing measurement error in a less reliable test. The tests' descriptive statistics are listed below, in Table 1. In addition to responses, Data Set A also included test takers' scale scores on a second test, the TOEIC Bridge, and information regarding the test takers' teachers for Listening classes. Data set B included information on the

individual classes each student was registered for. This additional information allowed for a Pearson correlation from scores to a second test and ANOVAs with categorical independent variables. It should be stressed that these analyses are essentially ad hoc, and done merely to examine differences between raw scores and IRT ability estimates in common statistical tests. Therefore, little attention will be given to the significance or meaning of the various results.

Table 5. Descriptive statistists of tests used.

	A: 2010 KSU Test	B: Western Music Test
k	100	100
N	654	234
Mean Score	48.50	56.40
SD	16.60	8.00
Reliability	0.91	0.75

Analyses using the 2010 KSU Test

The 2010 KSU Test is an older form of a placement test of English language listening and reading skills currently under piloting at the author's institution. In addition to test scores, the data set includes the TOEIC Bridge test scores of the students who took it, and information regarding each student's teacher. This permitted a) a Pearson correlation between the test and an external, validated measure of language proficiency, the TOEIC Bridge test, and b) an ANOVA of the effect of students' teachers on test scores.

In addition to raw scores on the test, ability estimates for the test under the Rasch model and the 2-Parameter Logistic Model were generated using the statistical software package JMP 8. A drawback of this study is that the JMP manual does not specify its estimation method, but it was observed that ability estimates produced for the Rasch model were identical to those produced by the program WINSTEPS, which uses JMLE.

The Pearson Correlation and ANOVA described above were then performed three times, each time using a different ability measure of the test as a variable: raw score, Rasch person measure, 2PL ability estimate. Analyses were then compared to determine if use of Rasch and 2PL ability estimates substantially altered the results of the experiments. Correlations between the raw scores of the test and Rasch and 2PL ability estimates are listed below. As Fan reported, correlations are very high, though very marginally lower between raw score and 2PL ability estimates.

Table 6. Correlations of raw scores of 2010 KSU Test to Rasch and 2PL IRT ability estimates.

	Raw Score
Rasch	0.997
2PL	0.986

Next, the test was correlated to the TOEIC Bridge test using the three scoring methods, as listed below in Table 3. The correlation for each is approximately 0.83; the difference between them is essentially indistinguishable.

Table 7. Correlations of 2010 KSU Test to the TOEIC Bridge Test by raw score and Rasch and 2PL IRT ability estimates

	TOEIC Bridge Test
Raw Score	0.833
Rasch	0.834
2PL	0.835

Next, an ANOVA was conducted on the effect of students' Listening class teachers on their scores, using the test's listening section scores calculated under all three methods. The F-Test was significant for each treatment (p = < .0001), though the R-Square and Adjusted R-Square for raw scores was slightly higher (marked in bold) than either IRT scoring method.

Table 8. One-way ANOVAs of differences on KSU Test scores by teacher using raw scores, Rasch measures, and 2PL ability estimates as dependent variables.

	Dependent: Raw Score	Dependent: Rasch Person Measure	Dependent: 2PL Ability Estimate
F Ratio	15.603	14.638	15.079
Prob > F	< .0001*	< .0001*	< .0001*
R-square	**0.241**	0.229	0.234
Adj R-square	**0.225**	0.214	0.219
Root Mean Square Error	7.396	0.772	0.884
Mean of Response	25.199	0.000 (person centered)	0.000 (person centered)
Observations (or Sum Wgts)	654	654	654

Analyses using the Western Music Test.

The Western Music Test was a less successful form of a low-stakes classroom test of students' listening comprehension of songs studied throughout a semester, and understanding of the expressions and vocabulary found in the lyrics. Reliability was low due to poorly targeted items and a low average point biserial correlation of approximately 0.19. Despite poor overall reliability no item point-biserial correlations were substantially negative, though two were effectively 0. Removing them did not appear to significantly improve split-half reliability.

The use of a test of lower reliability allows further examination of the theoretical advantages of ability estimates under the 2-parameter model. Although it does not pertain to the argument that IRT measures produce interval data, a potential advantage of the 2PL ability estimates is its use of item slopes in calculating ability. Under the Rasch model, items are assumed to have equal or approximately equal discrimination, meaning that equal weight is given to each question answered correctly. In contrast, the 2PL model weighs each item by its item discrimination, meaning successful endorsement of items with high discrimination contributes more to estimates of ability than items with low discrimination.

However, were the items of a test to fit the Rasch model and have roughly equal discrimination, any advantages offered by the 2PL ability estimate would become unobservable. This could be the case with the 2010 KSU Test, which was constructed in accordance to its ideal Test Information Function under the Rasch model, and uses items of fairly high (and relatively equal) discrimination for a multiple-choice test. Were 2PL ability estimates able to reduce measurement error, such an effect would likely be most measurable on a test with higher degrees of measurement error to begin with. A drawback of Fan's study was that descriptive statistics were not reported for the test analyzed, the Texas Assessment of Academic Skills (TAAS) for the 11th grade. Presumably, however, the State of Texas School Board employs tests of high reliability. If so, it could be argued that Fan's data set was not ideal for testing this possible advantage of 2PL ability estimates.

The test's raw scores and ability estimates under the Rasch and 2PL models were correlated, as can be seen in Table 5.

Table 9. Correlations between raw scores, Rasch person measures, and 2PL ability estimates of a test with a Cronbach alpha of 0.76.

	Raw Score	Rasch	2PL
Raw Score	1.000		
Rasch Person Measures	0.999	1.000	
2PL Ability Estimate	**0.935**	**0.942**	1.000

Correlations between raw scores and Rasch measures are very nearly 1. In this instance, however, the correlation between 2PL ability and raw scores is noticeably lower. What effect could this difference have on an experiment conducted using test scores as a variable? Although there is no data for a second test with which to correlate to, we can hypothesize how the test could correlate to another test of equivalent reliability under each ability measure by examining split-half correlations under each scoring method, as can be seen in Table 6.

Table 10. Split-half correlations for raw score and Rasch and 2PL ability estimates.

Raw Score	0.504
Rasch Ability Estimate	0.506
2PL	**0.571**

In this case, the difference in correlation is fairly sizeable, and would be enough to warrant the use of 2PL ability estimates for comparisons of test scores using Pearson correlation.

Finally, an ANOVA was conducted on test scores by class using each ability estimate method. Although differences in R-squared values were small, in this case there was a critical distinction: the ANOVA using the 2PL ability estimate had a p-value well below the critical threshold of 0.05,

and the other two ability estimates did not. In this instance, a technically statistically significant result was reached that would not have been had raw scores been used. Though the substantive difference remains negligible and <0.05 is a rather arbitrary value for "significance" (See Eidswick, this issue and Brown, this issue for further discussion), regrettably at many journals it can still make the difference between results that are considered publishable and results that are not.

Table 11. One-way ANOVAs of differences on test scores by class using raw scores, Rasch measures and 2PL ability estimates as dependent variables.

	Dependent: Raw Score	Dependent: Rasch Person Measure	Dependent: 2PL Ability Estimate
F Ratio	1.838	1.854	2.301
Prob > F	0.072	0.069	**0.022***
R-square	0.065	0.066	0.080
Adj R-square	0.030	0.030	0.045
Root Mean Square Error	8.651	0.466	0.978
Mean of Response	56.355	-0.003	-0.015
Observations (or Sum Wgts)	220	220	220

It appears that although differences are negligible for more reliable tests, for less reliable tests 2PL ability estimates can have greater efficacy than either raw scores or Rasch measures in statistical tests. To my knowledge this has not been documented in the literature. When told of it, experts in Item Response Theory have expressed surprise; Hambleton (personal communication) doubted that there could be much difference unless item point-biserials were negative. However, while this research marks my first formal study of the phenomenon, I have observed it in the past with a fair amount of consistency. It could be the case that researchers in psychometrics and educational assessment have a tendency to work with highly reliable tests that are less likely to reveal such differences to begin with. Unfortunately, however, testing instruments of poor reliability can still be quite common in other fields, and it is still possible for research using tests and surveys with reliability as low as 0.75 to be published in second language acquisition journals. Borsboom (2006) lamented that researchers in the social sciences often ignore advances in psychometrics and modern test theory. In a reply, Kane (2006) stated, "A great way to get their attention is to show them what you can do for them." Perhaps this application of 2PL ability estimates provides such an example.

Conclusions

In conclusion, raw scores and Rasch ability estimates are very highly correlated, and lead to extremely similar results when used in common statistical tests. For a well-constructed, reliable test, the same is true of 2PL ability estimates. However, in cases where the test has middling to poor reliability, 2PL ability estimates actually do appear to produce a somewhat more sensitive measure of a latent trait than raw scores, and their use as variables can result in meaningful differences in statistical tests.

I do not mean to diminish the usefulness of Rasch measurement, however. In closing, three things must be remembered:

1. *Rasch analysis should be seen as a method for examining and altering tests to reflect optimal measurement properties, not a magic transformation that automatically improves data.*

It should be noted that the first test examined had already been optimized in Winsteps, with items chosen to produce the optimal Test Information Function under the Rasch model. Misfitting items from pilot stages were not used in the final form. Therefore, although the final form worked just as well with raw scores as with Rasch measures, it had already benefitted from analysis under a Rasch framework; the analysis can be of benefit even if the raw scores of the resulting test are used.

2. *Rasch measures are of practical value even with near-perfect correlation to raw scores.*

Rasch measures provide a theoretical framework with which to examine interactions between test takers and individual items. They remain useful in equating scores between separate test forms, where different raw scores can result in identical measures. They can then be used as the basis for scale scores between forms. There is furthermore a persuasive theoretical argument that by accounting for intervals of difficulty between items, Rasch measures allow for comparable change scores, ensuring that, for example, the difference between the scores of 55 and 59 is indeed identical to the difference between the scores of 90 and 94.

3. *The 2PL ability estimates are only superior to raw scores if the test itself is inferior.*

While the 2PL ability estimates may be of value in salvaging data when a research instrument proves to have less than ideal reliability (and for whatever reason re-doing the experiment is no longer possible), it should be stressed that this technique should be viewed as a distant second to simply constructing a reliable test in the first place. As can be seen with the first data set, if the test is well constructed, results by all three scoring methods will be essentially identical.

References

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425-440, DOI: 10.1007/s11336-006-1447-6

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*, 357-381. DOI: 10.1177/0013164498058003001

Harwell, M. R. & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research, 71*(1), 105-131. DOI: 10.3102/00346543071001105

Linacre J.M. (1998). Do correlations prove scores linear? *Rasch Measurement Transactions, 12*(1), pp. 605-606. Retrieved January 16, 2012 from http://www.rasch.org/rmt/rmt121b.htm

MacDonald, P., & Paunonen, S.V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62.*

Rupp, A. & Zumbo, B. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When Pearson correlations are not enough. *Educational and Psychological Measurement, 64*(4), 588-599. DOI: 10.1177/0013164403261051

Wright, B. D. (1992). Raw scores are not linear measures: Rasch vs. Classical Test Theory CTT Comparison. *Rasch Measurement Transactions, 6*(1), 208. Retrieved January 16, 2012 from http://www.rasch.org/rmt/rmt61n.htm

The *akahon* publications: Their appeal and copyright concerns

Greg Wheeler
wheeler@sapmed.ac.jp
Sapporo Medical University

Abstract

The majority of those teaching in Japanese universities are likely familiar with the juku (often called "cram schools" in English) and their role in preparing students for the ideal of taking university entrance exams. Less well-known, perhaps, are the akahon (literally, "red book"), compilations of universities' past exams. When they do come up as a topic of discussion, they are often derided and the publisher of the akahon, Kyōgakusha, is frequently accused of violating copyright laws. Nevertheless, the akahon are extremely popular among hopeful university applicants. In the present short study, I outline what one can expect to find in typical akahon publications, discuss reasons for their appeal, and examine copyright concerns surrounding them.

Introduction

Although still a nerve-wracking experience for many, the prospect of sitting for a university's entrance exam does not always elicit the same amount of apprehension as it once did. One reason for this is that at an increasing number of Japanese universities, the exam is no longer the only means of admittance. A select number of students are able to enter universities every year via recommendations or through the Admission Office (AO) system. Additionally, at a time in which many universities' survival is contingent on greater student numbers, entrance exams at these schools are mostly formalities; acceptance is almost a foregone conclusion.

For those wishing to attend the more prestigious universities, however, admittance for most is still largely contingent on scores for those universities' entrance exams. Pressure on applicants to achieve high scores on these exams remains considerable, and thousands study at juku in order to better prepare themselves for the ordeal of taking these exams. Certainly, the juku are often criticized as having more interest in profit than in furthering the education of its students. Nevertheless, at a time in which public education is often perceived as lacking, the juku are seen by many students (and their parents) as necessary in order to achieve a good exam score.

Appeal of the akahon

Although the juku may be considered by many as the best way to achieve high exam scores, they are not inexpensive, and not everyone can afford to attend them. Alternatively (and in many cases, of course, additionally), students look for material that will aid them as they study on their own, of which there is no shortage available for purchase. Especially popular for those hoping to gain insight into the entrance exam of the university they wish to attend are the akahon, compilations of the university exams, published each year in May by Kyōgakusha. Easily recognizable by their bright red covers, the akahon are fixtures in major bookstores as well as being sold online.

Although the juku have gained a sort of international notoriety, much less is known about the akahon. In my own experience, when they come up as a topic of discussion among university faculty, both Japanese and foreign, they are usually ridiculed and held up as examples, along with

the textbooks published by the juku, of blatant copyright abuses. They are, however, extremely popular among university applicants, and there are several reasons why they are in high demand.

The most obvious appeal of the akahon is that they contain, almost entirely in their original form, the actual exams from universities. A typical akahon will include an individual university's entire exam from the past two or three years, although some will have up to as many as ten previous exams. There are also akahon that are compilations of sections of a university's exam, rather than the entire test. Someone interested in applying to Tokyo University, for example, who wishes to concentrate primarily on the English section of its entrance exam, can purchase an akahon for the university that contains the test's English section, often from as far back as 25 years.

In addition to the actual tests, every akahon includes answers to and detailed explanations of all questions appearing in the exams. It will also summarize recurring patterns or trends in exams and dispense advice regarding how an applicant should best prepare for them. Interestingly, there are frequent recommendations that students purchase various English textbooks published by companies other than Kyōgakusha. For the reputed higher-tier universities, full translations of the English passages that appear in exams are also usually provided.

At universities with competitive acceptance rates, one can find yearly figures for the number of applicants who take the exams, as well as the success rates. In some of the akahon, such data is provided for every department of the university. This can be of possible benefit to an applicant whose main objective is admittance to a certain university, with field of study not being of major concern. He or she can compare acceptance rates of each department and decide to apply to one which seemingly affords the greatest chance of admission.

There is also data, once again broken down by department when applicable, for mean test scores. Additionally, the lowest test scores from applicants who achieved admission are provided as well.

Concerns regarding actual usefulness and copyright

Overall, for hopeful test-takers the akahon are considered a useful and—in comparison to the juku—relatively inexpensive means of studying for entrance exams. (Prices vary depending on the number of exams included in each akahon, but most appear to cost approximately 2,000 yen.) On the other hand, there are issues of concern. The accuracy of a number of the answers they provide for exam questions, for example, is often questionable, as is that of the Japanese translations of passages that appear in the English sections of exams. Additionally, much of the advice provided to test-takers is based on trends noted from previous years, and seemingly assumes that a university's exam will follow the same basic format it has in the past. This is generally an accurate assessment; the majority of universities do not appear to make major changes in their exams from year to year. However, if a university's exam committee does make major revisions to its exam's format, those who have studied for it in a certain manner based on the advice from the akahon may encounter unexpected difficulties.

The greatest appeal of the akahon—their inclusion of actual university exams—also subjects them to frequent criticism and questions about possible copyright violations. It is common knowledge that much of the material that appears in universities' entrance exams, particularly the English sections, has been published elsewhere previously. Although Murphey (2005) and Wheeler (2009, 2011), to varying degree, have previously raised questions regarding the ethics of this, according to the basic premise of Article 36 from Japan's copyright laws, universities are allowed to use published material on their exams as long as profit is not their main goal. Kyōgakusha, however, is a private company, with profit being its primary objective. As such, the provisions detailed in

Article 36 regarding the reproduction of already existing material do not extend to the publisher. If Kyōgakusha wishes to use these works in the akahon, the publisher is obligated to receive consent from those holding copyright over them. Whether akahon publications comply fully with copyright laws has been a matter of question in the past. A lawsuit was filed in 2005, for example, by a group of Japanese authors who claimed that their works were being used in the akahon without their prior consent ("Nyūshi mondaishū," 2005).

Kyōgakusha posts on its akahon webpage (n.d.) that whenever possible, it establishes contact with copyright holders whose works they wish to publish, and receives permission to do so. Moreover, near the front of every akahon, there is an explanation of how to best use the akahon; included in this note is a message of gratitude to all authors who have granted the publisher permission to use their works. However, Kyōgakusha also notes on its webpage that there are instances in which no matter how exhaustive its search, it is unable to determine the authors of passages that appear in exams and displays a list, periodically updated, of passages from university exams in which the original authors or copyright holders are considered "unknown." One of its most recent posting (dated December 27, 2011) includes 209 passages, either reading or listening, from the exams of five universities (n.d.). All of these passages are from the English sections of these exams. This should not be entirely surprising; not all universities that utilize pre-existing material in their exams' English sections provide information concerning the original authors. In fact, from among the forty-four 2012 university exams I have observed, previously published material in the English section is included in 43, but references are provided in only 24. In the *kokugo* (i.e., Japanese literature and composition) sections of these exams, there are also passages that have been published previously, but in these cases, citations are provided for every passage.

Even if Kyōgakusha has indeed received authorization from the hundreds, or even thousands, of authors whose works it has identified as appearing in university exams, it could perhaps be more thorough in its attempts to identify the authors of passages whose names are not provided in the exams. Observing an earlier entry of its list of passages in which the publisher had been unable to discover the authors' identities, I was able to do so by simply copying one or two sentences of the passages in question onto Google's search engine. Kyōgakusha apparently had success in eventually obtaining information on these authors as well; in its most recent entry regarding unknown authors, these passages no longer appear. However, the fact remains that in the previous year's akahon, these passages were published without providing any information about the authors, and presumably without their prior consent.

Universities, of course, could make matters easier for Kyōgakusha by citing on their exams the authors whose works they have used. However, it is almost certainly the responsibility of the publisher to check any material for sources. This could seemingly be obtained fairly easily, simply by requesting the information from the university in question. At the time of this writing, for example, Obunsha, another publisher of entrance exam compilations, requested detailed information from my university regarding the authors whose works appeared in the exam. It does appear, based on information gathered from an informal survey of a number of universities in Hokkaido, that Kyōgakusha makes similar requests on a number of occasions. In the case of at least one university, in fact, citations of authors whose works were used in the university's entrance exam appear in the akahon for that school, but were not on the actual test itself; they were added by Kyōgakusha prior to publication. However, unless the lists of passages of unknown authors on its website are from the exams of universities that refused to disclose information regarding the sources of these passages, Kyōgakusha could seemingly be more consistent in reaching out to the universities.

Relationship with universities

Although the akahon are compilations of university exams, Kyōgakusha's connections with the universities whose exams appear in its publications appear to vary. It is often said that rather than approach universities directly, Kyōgakusha and other publishers of entrance exam compilations procure copies of an exam by purchasing it from a test-taker upon the exam's completion. However, there is also indication that at least a few universities may be affiliated with the akahon. In the explanation note that appears in each akahon (mentioned above), in addition to thanking the authors from whom the publisher received authorization to use their works, Kyōgakusha also expresses gratitude to "those connected to the universities who provided material [資料のご提供をいただいた大学関係者各位]."

The university at which I teach currently has no official relationship with Kyōgakusha, and receives no compensation from the publisher for its use of the university exam. A university that does enter into a monetary relationship with Kyōgakusha, however, would be advised to exercise caution. According to the Copyright Research and Information Center (CRIC), if the university's exam includes pre-existing material, its right to authorize publishers to use said exam is limited (Copyright case study, n.d.). CRIC posits that the university can only "give authorization to the publisher under the condition that the publisher obtains necessary authorization from all the relevant copyright owners (n.d.)."

Copyright issues aside, and here Kyōgakusha appears to be far more compliant regarding this matter than many assume, the akahon, while not perfect, do offer prospective university students a means of studying for entrance exams that is far cheaper than that of attending the juku. If nothing else, simply that they afford test-takers the opportunity to study from actual past exams is of considerable appeal, and does much to explain their popularity.

References

Akahon.net (n.d.). *Chosakusha no kata wo sagashiteimasu* [Searching for authors]. Retrieved from http://akahon.net/settlement/

Copyright Research and Information Center (n.d.). *Copyright case study vol. 1: Formal education and copyright*. Retrieved from http://www.cric.or.jp/cric_e/cs_1/case1_qa.html#q7

Murphey, T. (2005). Entrance exams breaking copyright law? Academically unethical? *Shiken: JALT Testing & Evaluation SIG Newsletter, 9*(2), 23-28.

Wheeler, G. (2009). Copyright issues concerning Japan's university entrance exams. *The Language Teacher, 33*(8), 3-7.

Wheeler, G. (2011). Addressing copyright concerns regarding Japanese university entrance exams. In A. Stewart (Ed.), *JALT2010 Conference Proceedings*. Tokyo: JALT.

"Nyūshi mondaishū ni mudan tensai" sakkara, "akahon" shuppan moto ni baishō motome teiso [Reproducing entrance examination problems without release. Authors file lawsuit for damages against "redbooks"]. (2005, April 28). *Yomiuri Shimbun,* p. 38.

Questions and answers about language testing statistics:

What do distributions, assumptions, significance vs. meaningfulness, multiple statistical tests, causality, and null results have in common?

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

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. 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.

Distributions underlie everything else

Statistical studies investigate variables, and those variables are operationalized (i.e., observed and quantified) into scales that are nominal, ordinal, interval, or ratio (for definitions and examples of these different types of scales, see Brown, 2011a). The variables of focus in the majority of language studies are observed or measured as interval or ratio scales (known collectively as continuous scales). For many statistical analyses, such continuous scales need to be normally distributed, or if they are not, the researcher needs to consider what the effect might be of that lack of normality.

As I will explain in the next section, most statistical analysis make certain assumptions, the first of which in many cases is the assumption of normality (i.e., for the statistics to work well, the distributions in the continuous scales must be normal, or approximately normal. This is particularly important for correlational statistics, or statistics that involve correlation in any way (e.g., reliability estimates, regression analysis, factor analysis, structural equation modeling, analysis of covariance, etc.). To insure that their statistics can function appropriately, researchers always need to check the assumptions that underlie those statistics. Sadly, that is not often the case in second language research. At very least, researchers should provide descriptive statistics (including means, standard deviations, minimum and maximum values, numbers of people and items, reliability estimates, etc.) so that readers can examine for themselves the degree to which important assumptions like normality, equality of variances, reliability, and so forth have been met. That is why I report descriptive statistics and reliability estimates in my own studies before I do anything else. If all quantitative researchers would do the same, that habit would go a long way

toward increasing the quality and interpretability of the quantitative research in our field because the distributions of data (normal or otherwise) underlie everything else in statistical analyses.

Assumptions must be examined

Why do statistical tests have assumptions? The various statistical tests that researchers use were all created and tested for application under certain conditions, and they were found to work under those conditions. If those conditions do not obtain, that is, if the assumptions are not met, researchers cannot be sure if their statistics are being properly applied and accurately doing what they were designed to do. For example, the common Pearson product-moment correlation coefficient assumes that (a) the data for both variables are on a continuous scale, (b) the observations within those scales are independent of each other, (c) the distributions for the scales are normal, and (d) the relationship between the two scales is linear (for explanations of how these assumptions are defined, how they can be checked, and how the results should be interpreted when violations of the assumptions occur, see Brown, 2001, pp. 140-143). If the assumptions are met, all is well, and the researcher can interpret the results within the limits of probability that the statistics indicate. However, if the assumptions are not met, the researcher cannot be sure of the interpretations. For example, in the case of the correlation coefficient, if the distribution for one of the scales (or both) is skewed (i.e., non-normal with values scrunched up at one or the other end of the scale), it may not be appropriate to use a correlation coefficient at all, or it may be wise to adjust for the violation of the assumption by normalizing the variables. Alternatively, it may be necessary to interpret the resulting correlation coefficient very cautiously, while recognizing the likely effects of the skewing. In my experience, the likely effect when one (or both) variables is skewed is that the magnitude of any resulting correlation coefficient will tend to be depressed (i.e., will tend to provide an underestimate of the actual state of affairs). In any case, ignoring the assumptions of the seemingly simple correlation coefficient is ill-advised.

I don't want to get down in the weeds here by discussing the assumptions of every statistical procedure. The point is that for virtually every form of statistical analysis, two things are true: there is a standard error for that statistic (see Brown, 2011b), and there are assumptions that should be considered in setting up, conducting, and interpreting the analysis of that statistic (for an overview of the assumptions underlying a wide variety of statistical analyses, see Brown, 1992).

Statistical significance does not assure meaningfulness

One of the biggest problems in second language quantitative research occurs when researchers treat statistical significance as though it indicates meaningfulness. I have spent 35 years chanting that statistical significance and meaningfulness are different things, yet nothing seems to change. It is a fact that a study with a sufficiently large sample size can produce statistics (e.g., correlation coefficients, t-tests, etc.) that are statistically significant for even small degrees of relationship or small mean differences. Those p-values that lead to interpretations of statistical significance (e.g., $p < .05$ for a particular correlation coefficient) only reveal the probability that the statistic occurred by chance alone (e.g., $p < .05$ for a correlation coefficient means that there is only a 5% chance that correlation coefficient of this magnitude would occur by chance alone). That p-value does not mean that the correlation or mean difference or whatever is being tested is large, interesting, noteworthy, or meaningful. These characteristics can only be determined by looking at things like the magnitude of the correlation within the particular research context or the size of the mean difference in the context. For instance, it is perfectly valid to ask if a significant (with p

< .01) correlation of .40 found in a particular study is also meaningful and interesting. But the researcher cannot answer that question without considering the magnitude of the statistical results within the context of the specific research situation. Sometimes, a small correlation is very interesting because the researcher is looking for any sign of relationship. In such a situation, .40 would be meaningful. Other times (e.g., when costs or other stakes are very high), only a strong correlation of say .90 or higher will be meaningful. Similarly, a mean difference of 10 points on a 20 point scale might seem very interesting, but on a 1000 point scale 10 points might be far from interesting, especially if it took 300 hours of instruction to produce that one percent difference. So clearly, interpreting the meaningfulness of any statistic is different from, and additional to, first deciding whether that result has a high probability of being a non-chance statistical finding. In other words, while significance is a precondition for interpreting a statistic result at all (after all nobody wants to interpret a result that is due to chance alone), the degree to which the same statistic is interesting or meaningful will depend on the magnitude of the results and the context in which they were found. That is why statistical significance, though a precondition for meaningfulness, does not assure meaningfulness.

Multiple statistical tests cloud interpretations

Multiple statistical tests are another big problem in our research that my chanting does not seem to have affected. This phenomenon occurs when researchers perform multiple statistical tests without adjusting their p-values for that fact. During the last 35 years, I have observed multiple statistical tests in so many second language research studies that I can't even guess how many there are out there. Yet, I continue to staunchly believe (because of my training and experience with statistics) that multiple statistical tests create important problems in interpreting statistical results. I have explained this issue elsewhere in more detail (e.g., Brown, 1990, 2001, pp. 169-171, 2008), and I am not alone in holding this view (e.g., Dayton, 1970, pp. 37-49; Kirk, 1968, pp. 69-98; Shavelson, 1981, pp. 447-448; and so forth).

In brief, the problem is that conducting multiple statistical tests seriously clouds the interpretation of resulting statistical tests, usually by increasing the probability of finding spuriously significant results (i.e., results that are not really significant, popularly known as "false positives"). This problem is amplified by the fact that researchers who produce spuriously significant results do not know which of their results are spuriously significant, so even results that might actually be significant cannot be trusted. The kindest way to put this problem is that multiple statistical tests cloud interpretations. Sadly, with proper use of the analysis of variance (ANOVA) family of statistics, the effects of such multiple comparisons can be controlled (by including all of the comparisons in one omnibus ANOVA design) or minimized (by using the Bonferroni adjustments when multiple comparisons cannot be avoided) [For more on the latter topic, see Brown, 2001, pp. 169-171, 2008].

Causal interpretations are risky

Another axiom that I live by is that it is irresponsible to interpret significant statistics, even ones that appear to be meaningful, especially correlation coefficients, as indicating causality. Just because two sets of numbers seem to be related does not mean that either variable is causing the other. There are many reasons for two sets of numbers to be correlated without either causing the other. Most notably a third factor may be causing both of the variables of interest to be related. For example, when I was young and stupid, I smoked and drank coffee like my life depended on it. In fact, the numbers of cigarettes per hour and the number of cups of coffee per hour were

probably significantly correlated (at say $p < .01$). Does that mean that the coffee was causing the cigarettes or vice versa? No, of course not. There was simply a relationship. A third variable was probably causing both (e.g., fatigue, or need for stimulation, or social pressures, or advertising, or some combination of these factors). The message should be clear: be very careful if you are tempted to interpret causation based on any statistic. There may always be an alternative explanation that you overlooked for your result. That is why causal interpretations are so risky.

Null results do not mean sameness

Researchers are often tempted to interpret a lack of statistical significance (e.g., the probability is greater than 5%, or $p > .05$) as showing statistical sameness. For example, a researcher may use two ESL classes as experimental groups with one group getting some specific instructional treatment and the other group serving as a control group that gets some unrelated "placebo" treatment. Since the two groups were samples of convenience (i.e., not randomly assigned), the teacher/researcher will be tempted to compare the two groups on some form of pretest to see if they are the same at the beginning of the experiment. Naturally, they are never exactly the same, so the researcher performs a t-test to see if the difference is significant and infers (or counts on the reader to infer) from a non-significant result (i.e., $p > .05$) that the two groups were therefore statistically the same at the beginning of the study. This is not a correct inference, that is, the $p > .05$ does not indicate the probability that the two groups were the same on average. It does indicate that the researcher was unable to establish that the mean difference was statistically significant. Such a result can easily occur simply because the research design lacked sufficient power to detect a statistically significant result. Many factors can contribute to a lack of power: a sample size that is too small, measurement that lacks reliability, limited variation in ability levels for the construct being measured, etc. To determine if this is the case, procedures known as power analysis need to be included to defend any conclusion about the probability of sameness for the means of two groups. The bottom line is that a finding of no statistically significant mean difference indicates that the study was unable to establish significance, not that the two means are the same. [For further explanation of this issue, see Brown (2007a; 2007b).]

Conclusion

In the title of this column, I asked the following question: What do distributions, assumptions, significance vs. meaningfulness, multiple statistical tests, causality, and null results have in common? The simple answer is that these are six of my pet statistical peeves. To recap briefly, my pet statistical peeves are that researchers in our field often:

1. Forget to consider the potential effects of their data distributions on their statistical results (and foolishly forget to report descriptive statistics)
2. Fail to check the assumptions for the statistics they use, much less consider what violations of those assumptions mean for the interpretation of their results
3. Act as if statistical significance means that the results of their study are interesting and meaningful, which is flat out not true
4. Let multiple statistical tests cloud their interpretations
5. Make unjustified causal interpretations of their results
6. And, treat non-significant results as though they indicate the sameness of two groups

Why should anyone care about my pet statistical peeves? These peeves have developed over 35 years of experience in the ESL/EFL/Applied Linguistics field, and they are based on reading thousands of statistical studies in which I have witnessed researchers overinterpreting, underinterpreting, and/or misinterpreting their statistical results because the researchers were either ignorant of these six sets of issues or willfully ignored them. More importantly such overinterpretation, underinterpretation, and/or misinterpretation of statistical results means that the interpretations were wrong in important ways. And yet, they serve as the knowledge base of our field.

In direct answer to your question, the six sets of issues covered in this column serve as principals that are worth knowing because they are important to the quality of the statistical research in our field and because they "apply across the board to many types of statistical analyses." As a consumer of statistical studies, you can help improve the quality of the research in our field by paying attention to these issues whenever you pick up a professional journal and read quantitative research studies. My guess is that you already read such studies critically in terms of their content, but you might now want to also read them critically in terms of their statistical research methods. You can help increase the quality of the quantitative research in our field by being a critical reader, by spreading the word about these problems to your colleagues, and by complaining in letters to the editors of professional journals where you see researchers ignore these six sets of issues. Together we can help improve the statistical research methods used in the research of our field by refusing to tolerate shoddy work. How can that help but be good for the field, and good for our knowledge about second language learning and teaching?

References

Brown, J. D. (1990). The use of multiple t-tests in language research. *TESOL Quarterly, 24*(4), 770-773.

Brown, J. D. (1992). Statistics as a foreign language—Part 2: More things to look for in reading statistical language studies. *TESOL Quarterly, 26*(4), 629-664.

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

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. Also retrieved from the World Wide Web at http://www.jalt.org/test/bro_25.htm

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. Also retrieved from the World Wide Web at http://www.jalt.org/test/bro_26.htm

Brown, J. D. (2008). Statistics Corner. Questions and answers about language testing statistics: The Bonferroni adjustment. *Shiken: JALT Testing & Evaluation SIG Newsletter, 12*(1), 23-28. Also retrieved from the World Wide Web at http://www.jalt.org/test/bro_27.htm

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. Also retrieved from the World Wide Web at http://www.jalt.org/test/bro_34.htm

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. Also retrieved from the World Wide Web at http://www.jalt.org/test/bro_35.htm

Dayton, C. M. (1970). *The design of educational experiments*. New York: McGraw-Hill.

Kirk, R. E. (1968). Experimental design: Procedures for the behavioral sciences. Belmont, CA: Brooks/Cole.

Shavelson, R. J. (1981). Statistical reasoning for the behavioral sciences. Boston: Allyn & Bacon.

Where to Submit Questions:

Please submit questions for this column to the following e-mail or snail-mail addresses:

brownj@hawaii.edu.

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

Your question can remain anonymous if you so desire.

jMetrik 2.1

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

Many researchers are curious about Rasch analysis and would like to try it with their own data, and most have a need for classical test statistics from time to time. However, with prices ranging from $150 to well over $1000 (US), test software can be a major investment, leaving researchers unsure of where to get started. In our first software column, we will introduce free alternative to commercial packages called jMetrik, and detail how to get started with the program.

What is jMetrik?

jMetrik is a free, open-source software tool for psychometrics. Unlike packages for R which rely on command lines, it offers a Graphical User Interface, making it easy for beginners to navigate. It is Java-based, so it is 100% cross-platform, offering identical operation on Windows, Macintosh, and Linux. It offers a range of statistical and psychometric functions that are usually not available for free, including:

- Descriptive statistics
- Classical test theory (CTT) item analyses
- Rasch modeling
- Test equating
- DIF analyses
- Graphing
- Some confirmatory factor analysis functions

Although it can be a little rough around the edges, it is a convenient tool for testing on a budget. In this, the first Software Corner article, I will explain the basics of getting started with jMetrik.

Obtaining and installing the software

The software is freely available from the following URL:

http://www.itemanalysis.com/user-form.php

Enter your name, email, and location, and you will be directed to a page from which you can download the version for your operating system. Note that if you are a Mac user, you may need to install Java separately. The page includes a link to do so.

Formatting and importing your data

Probably the most confusing aspect of getting started with jMetrik is simply entering your data, as jMetrik, like RUMM2020, follows a database model rather than a file-based model. This is very convenient for keeping all of your data sets and results together for a project, but comes at the cost of a somewhat more-involved data importing process initially.

Laying out your data

In all likelihood, your data is in an Excel (or compatible) spreadsheet. Although jMetrik cannot read Excel files, the data can be easily exported to a text-based format that can be read. In this section I will explain the process for tab-delimited files (although jMetrik supports other text formats as well).

1. Lay out your data.

In Excel, lay out your data so that cases are in rows and items are in columns. You may also include non-item variables such as case IDs and gender (see Fig. 1). **Be sure to save this in Excel before moving on to the next step.**

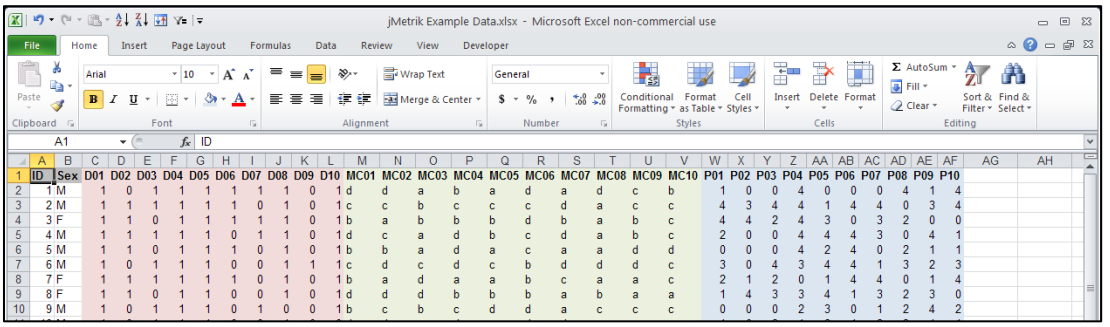

Figure 2. **Example data layout in Microsoft Excel, with case ID, sex, dichotomous items, multiple-choice items, and polytomous items.**

2. Save your data as tab-delimited text.

Go to the File tab (Windows) or menu (Macintosh), and choose "Save as." In the save dialog, set the format to tab-delimited text. See Figures 1 (Windows) and 2 (Macintosh) for the locations of this option.

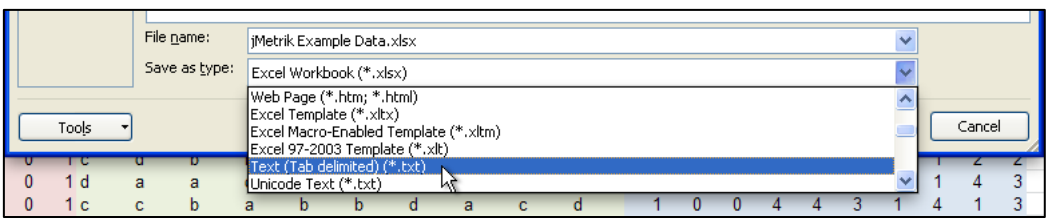

Figure 3. **Saving as tab-delimited text in Excel for Windows.**

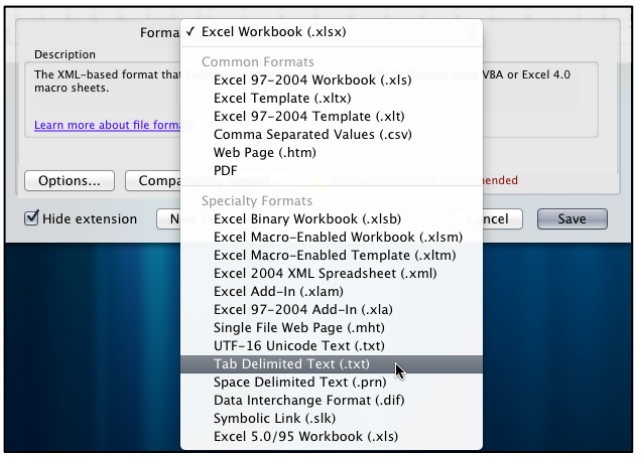

Figure 4. **Saving as tab-delimited test in Excel for Macintosh.**

You will see two dialog boxes asking if you would like save only the active sheet, and then if you would like to continue to save and lose formatting. Click "OK" and "Yes" to both of these. They will not affect your Excel data, which you saved before exporting a plain-text copy.

Importing your data into jMetrik

You cannot import your data into jMetrik until you have defined a database. Before doing that, however, you may want to create a new workspace, which can hold multiple databases and outputs for a project. This is entirely optional.

1. (Optional) Add a new workspace.

Workspaces allow you to keep projects totally separate. If you will be using jMetrik for several projects at the same time, this step is recommended. When the software starts, it uses a default workspace. If you would like a separate one, choose "Add Workspace…" from the Manage menu. Give it a name and a location to save it, and click "Add" (see Figure 4).

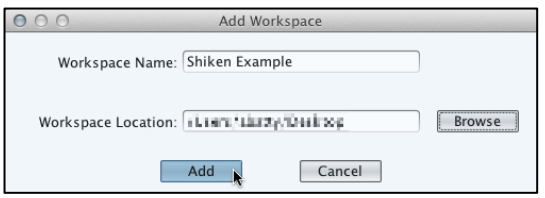

Figure 5. Adding a workspace.

Once you create a new workspace, you will need to change to it by going to the Manage menu and choosing "Change Workspace…" You will be presented with a list of available workspaces. Choose your new workspace, as in Figure 5.

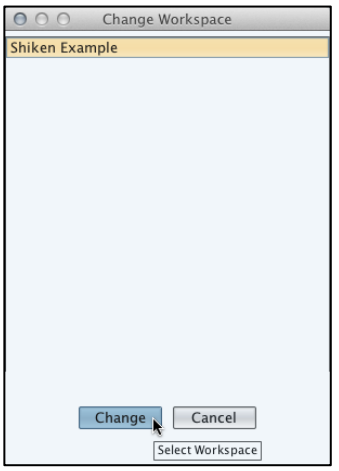

Figure 6. Changing to the new workspace.

2. Add a new database to the workspace.

After either creating a new workspace or electing to simply use the default, go to the Manage menu and select "New Database…" Give the new database a descriptive name. This database will hold all the data associated with your project, so it should be something easy to identify.

3. Import your data table.

The database created in the last step is empty. The next step is to add a data table. Go to the Manage window and select "Import Data…" In the open dialog that appears, navigate to the data file you exported from Excel and set the delimiter to "Tab". If your data file has the variable names in the first row, be certain to select that option. Click "Browse" to import the data. See Figure 6.

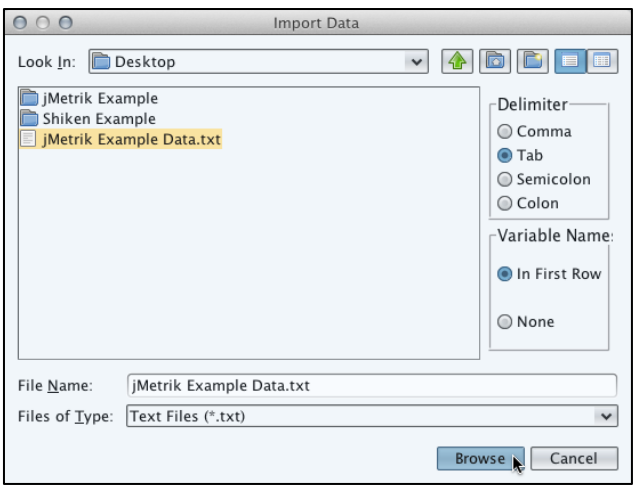

Figure 7. Import data dialog box.

In the dialog that appears, give the data set a name. You cannot use spaces in this name. Leave "Rows to Scan" blank to include all data. Finally, click "Import." See Figure 7 for an example.

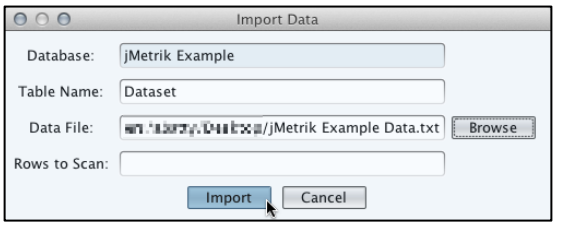

Figure 8. Creating the data table.

Scoring Item Responses

Because jMetrik can accommodate many kinds of item data, you will need to provide the scoring information for each item (anything that is not scored will be listed as "Not Item." Although this can be a little tedious with multiple-choice (MC) data, it does allow for CTT distractor analyses. To begin this process, click the "Variables" tab at the bottom of the jMetrik window (see Figure 8).

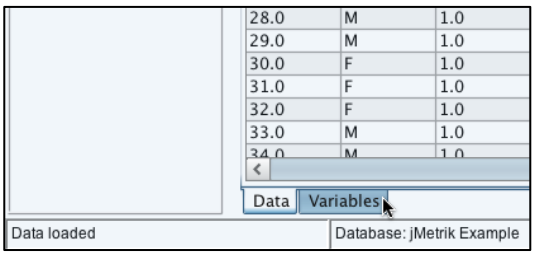

Figure 9. Switching to the variable tab.

In the variable tab, click the cell in the "Scoring" column for the first item. The Option Scoring dialog will appear. Refer to the following sections for binary, MC, and polytomous data.

Scoring binary items

1. Set the scoring for the first item.

In the Option Scoring dialog, enter "0" in the Option column, and "0" in the Score column. On the next line, insert a "1" in both fields. **Be sure to move to the next line or the values will not be registered.** See Figure 9.

2. Replicate the scoring for the rest of the binary items.

After entering the scoring for the first item, click "Replicate," as in Figure 9.

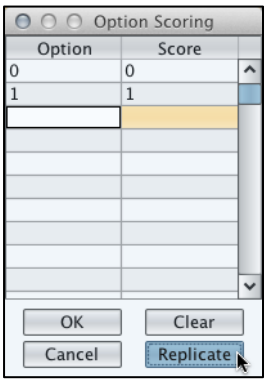

Figure 10. Option Scoring for binary items.

3. Set the items to which to replicate the first item's scoring.

After clicking "Replicate," a dialog box will appear. On the left are all of the variables in the data table. Select the items to which you wish to apply the first item's scoring on the left (NOTE: Click the first item and then shift-click the last item to select a range.) and click the right arrow to add it to the list to which the scoring will be added. See Figure 10.

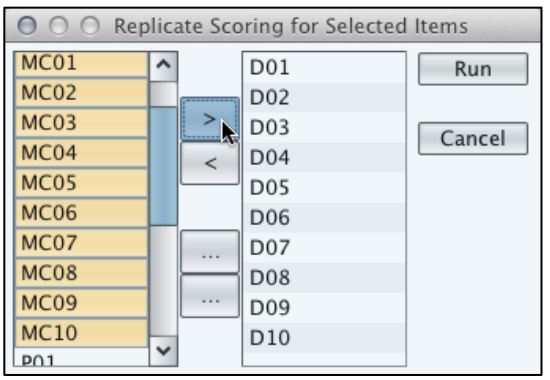

Figure 11. **Setting items to which to replicate scoring.**

4. Finish replication.

Click "Run" on the replication dialog, and "Okay" on the verification.

Scoring MC items

MC items are a little more difficult, since there are multiple scoring possibilities, depending on the correct option for the item in question.

1. Set the scoring for the first item.

Once again, click in the Scoring column for an item and enter the option scoring for that item (see Figure 11).

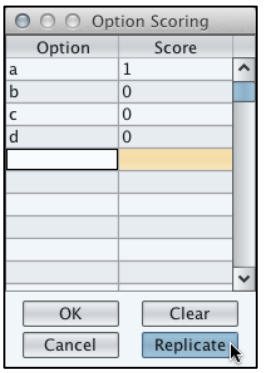

Figure 12. Option Scoring for MC items.

2. Replicate the scoring for the rest of the items with the same scoring.

Replicate the scoring for the first item you scored to every item with the same correct answer. For example, all the items for which "a" is the correct answer should have the scoring showing in Figure 11 replicated to them. Repeat for "b", "c", "d", etc. In the Replicate dialog, you can select multiple items with ctrl-click (Windows) or cmd-click (Macintosh).

Scoring polytomous items

Since jMetrik includes the rating scale and partial credit models, polytomous data is a possibility. Scoring of this data follows the same logic as that of the binary data. Simply set each number to its value. NOTE: If you need to collapse categories, simply assign the same score to two categories. Refer to Figure 12 for an example. Replicate for all the items to which this scoring is relevant.

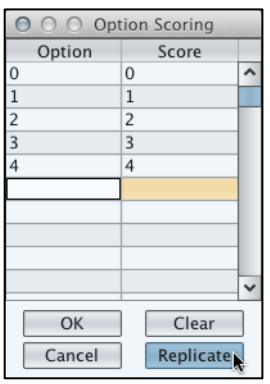

Figure 13. **Option Scoring for polytomous items.**

Finishing variable definition

On the Variables tab, you may add easy-to-read labels to the variables if you wish.

Analyses

Once the data are imported and scored, analyses are fairly straightforward with the graphical user interface. Overall, the interface for these analyses is very similar to that of SPSS, and to the

Replicate dialogs discussed above. Analyses and graphs output to tabs in the jMetrik window. Refer to the following sections for saving this information for use in other analyses or elsewhere.

Saving Rasch scores to the data table

Some analyses, particularly Rasch, can write the results to variables in the data table, to be used in further analyses. In the case of Rasch analyses, this option is located under the Person tab in the Rasch analysis dialog (see Figure 13).

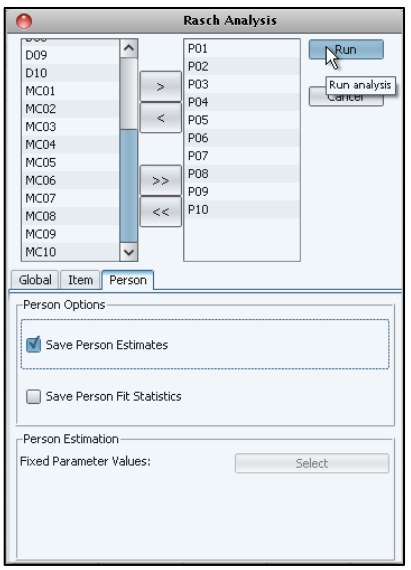

Figure 14. **Saving Rasch person estimates to the data table.**

Saving text outputs

If your analysis has created a text output that you would like to use elsewhere, click the "Text File Save As" button in the toolbar, or select "Text File Save As" from the File menu (see Figure 14).

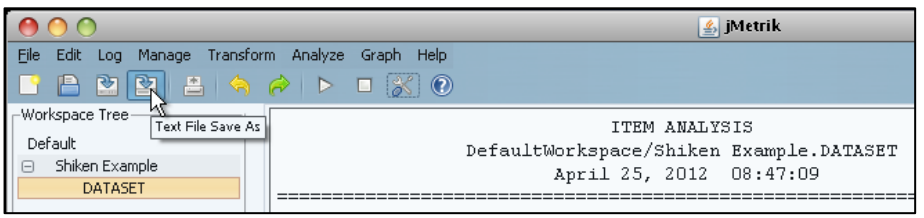

Figure 15. **The Text File Save As button in the toolbar.**

Saving graphical outputs

Graphs can be saved by right-clicking them and selecting "Save as…" The file that is created is in the Portable Network Graphics (.png) format. See Figure 15.

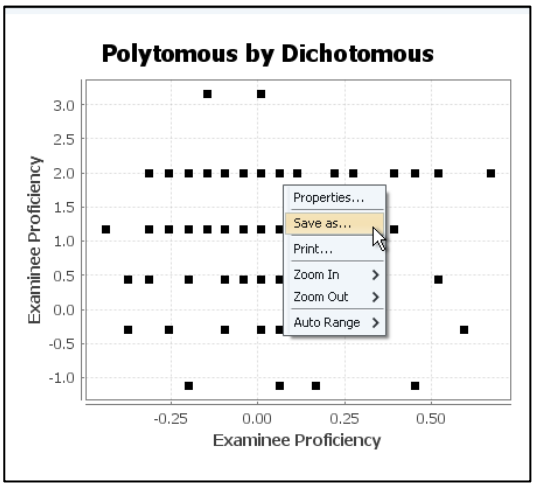

Figure 16. Saving a graph to a file.

Conclusion

Although setting up the data for analysis can be tedious, jMetrik offers researchers and students on a budget a suite of psychometric analyses for free. In a field where even software with very limited functionality is routinely priced in the hundreds of US dollars, this is a welcome alternative. Despite having an ever-growing number of psychometric software tools at my disposal, the lack of licensing difficulties and cross-platform compatibility of jMetrik has made it my first go-to tool for quick or exploratory analyses. I encourage you to try it out; I think you will be surprised by what it can offer for the price.

Upcoming Language Testing Events

The 11th Annual JALT Pan-SIG Conference: June 16 – 17, 2012

Abstract submissions: (closed)
Venue: Hiroshima University, Hiroshima
Conference homepage: http://www.pansig.org/2012/
—Featuring a joint workshop by the TEval and Framework and Language Portfolio SIGs!

Pacific Rim Objective Measurement Symposium (PROMS): August 6 – 9, 2012

Abstract submissions: (closed)
Venue: Jiaxing University, Jiaxing, China
Conference homepage: http://cfs.zjxu.edu.cn/proms/index.htm
—Featuring presentations by several JALT TEval SIG members!

The 16th JLTA Annual Conference (JLTA 2012): October 27, 2012

Abstract submissions: June 1 – July 8
Venue: Senshu University (Ikuta Campus), Kawasaki, Kanagawa
Conference homepage: http://jlta.ac/

Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ) First Annual Conference: November 9-10, 2012

Abstract submissions: (closed)
Venue: University of Sydney, Sydney, Australia
Conference homepage: http://www.altaanz.org/altaanz-conferences.html

Shiken Research Bulletin Editorial Board

Submissions

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

jaltteval+srb@gmail.com