'real' evidence of the existence of problems. This, I believe, would be seen as a more constructive form of criticism and would have a far greater chance of reaching those test makers referred to above.

While the thrust of this reply has taken a rather negative view of the Brown and Yamashita article, it is not meant merely to criticize what is a valuable and solid first step in the process of evaluating Japanese university entrance tests. In opening a debate on the reliability and validity of these examinations the article has confronted an issue of growing importance, and has raised a series of questions which researchers should now strive to answer with empirical evidence. These questions include:

Is there evidence of a topic awareness bias in some tests?
How harmful is the dependence on translation?
Can we establish the content and construct validity of these tests?

The Authors Respond to O'Sullivan's Letter to *JALT Journal*: Out Of Criticism Comes Knowledge

James Dean Brown
University Of Hawaii At Manoa

Sayoko Okada Yamashita
International Christian University

We would like to begin by thanking Barry O'Sullivan for his criticisms of Brown and Yamashita (1995a), as well as for his words of praise.

Taking the criticisms first, as far as we can tell, his primary complaints are that there are "quite serious problems" with our study in that:

1. "the design of the study severely reduces the possibility of using the data,"
2. we do not provide enough "detail and 'history'," and

3. we provide "no empirical evidence of problems of validity and reliability in any of the tests . . ."

Beginning with the issue of design, we purposely chose to use a descriptive approach rather than an inferential one because of well-justified concerns about the types and number of statistical comparisons that would have been necessary in such a statistical study (for more on this topic, see Brown, 1988). We also chose the descriptive route out of consideration for the audience of the *JALT Journal*, who are by-and-large hard-working teachers with little or no training in advanced statistics.

With regard to the issue of not providing enough detail, the amount of data involved in such a study necessarily involves making decisions along the way about what to include and what to exclude. We did this to the best of our abilities providing a tremendous amount of detail in a very limited space, but apparently, what we did was not up to Mr. O'Sullivan's expectations.

As for the issue of providing "history", we certainly looked for such "history" in the literature and found nothing. That is why we did our study, that is why we set out to provide base-line data, and that is why we have begun to create "history" by studying the same examinations in subsequent years. For instance, Brown and Yamashita (1995b) compares the 1994 tests to the 1993 tests described in Brown and Yamashita (1995a).

As for failing to provide evidence of the lack of reliability and validity of the tests, it is primarily the responsibility of the test developers (not the general public or the teaching profession or Brown and Yamashita) to provide evidence of the reliability and validity of the tests. As the American Psychological Association has put it (CDSEPT, 1985), "Typically, test developers and publishers have primary responsibility for obtaining and reporting evidence concerning reliability and errors of measurement adequate for the intended uses" (p. 19). They also state that "evidence of validity should be presented for the major types of inferences for which the use of a test is recommended" (p. 13). To our knowledge, no such evidence exists for the university entrance examinations in Japan. In addition, when we have requested such information from a number of universities and/or sought access to the data in order to study these issues ourselves, we have encountered resistance, secrecy, and a total lack of cooperation. A black hole of information exists about these important examinations from which no light can escape. Hence, we can only conclude, as we did in Brown and Yamashita (1995a & 1995b), that problems may exist with the reliability and valid-

ity of these tests. Naturally, we would welcome studies of these issues and would ourselves happily participate.

We would like to emphasize the fact that Mr. O'Sullivan was not entirely negative about our study. For instance, he stated that (a) our study "serves to highlight the lack of published accounts of empirical research in the area of university entrance test evaluation in Japan," (b) our paper provides "a valuable and solid first step in the process of evaluating Japanese university entrance tests", and (c) "in opening the debate on the reliability and validity of these examinations, the article has confronted an issue of growing importance,..."

He ends by calling for "empirical evidence" that addresses three questions:

1. "Is there evidence of a topic awareness bias in sometests?"
2. "How harmful is the dependence on translation?"
3. "Can we establish the content and construct validity of these tests?"

We would like to end by seconding his call for further research and adding to his list a number of other questions that occurred to us along the way:

4. How are norms established on these tests, and how do they vary from university to university and year to year?
5. What evidence is there for the reliability of these university entrance examinations (e.g., what is the K-R20, or Cronbach alpha reliability of these tests)?
6. What evidence is there for the decision reliability of these exams (i.e., what is the standard error of measurement, and how is it used to make university admissions decisions responsible and fair)?
7. What evidence is there for the content, construct, criterion-related, face, decision, or social validity of these tests (for more on these types of validity, see Brown, 1995a or 1995b)?
8. How are standards set for the cut-points used in deciding who will be admitted and who will not? Are state mastery methods used? Or, test-centered continuum methods? Or, student-centered continuum methods? Are rational methods used at all? (for more on standards setting, see Brown, 1995b)
9. Why do the examinations cost so much given the relatively cheap and easy-to-score formats that are used? Or put another way, why is it that communicative listening and speak-

ing subtests are not used on these exams even though there is apparently plenty of revenue to support such sound testing practices?

10. What is the impact of the "washback" effect of these tests on the educational system? In particular, what is their effect on the teaching of English?

The very fact that Mr. O'Sullivan felt compelled to react to our study is an encouraging sign. We would like to challenge him and any other readers who are interested in this issue to do their own research on the university entrance examinations so that all of us can begin to understand and perhaps ameliorate any existing negative effects of the "examination hell" that hundreds of thousands of students in all corners of Japan face year after year after year.

References

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

Brown, J.D. (1995a). *The elements of language curriculum: A systematic approach to program development.* New York: Heinle & Heinle Publishers.

Brown, J.D. (1995b). *Testing in language programs.* Englewood Cliffs, NJ: Prentice-Hall Publishers.

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

Brown, J.D. & S.O. Yamashita. (1995b). English language entrance examinations at Japanese universities: 1993 and 1994. In J.D. Brown and S.O. Yamashita (Eds.) *Language Testing in Japan.* Tokyo: Japanese Association for Language Teaching.

CDSEPT (Committee to Develop Standards for Educational and Psychological Testing). (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.