Building arguments to help decide ethical questions in language testing: Language testing SIG panel discussion

Randy Thrasher
Okinawa Christian University

Reference data:
Thrasher, R. (2006). Building arguments to help decide ethical questions in language testing: Language Testing SIG panel discussion. In K. Bradford-Watts, C. Ikeguchi, & M. Swanson (Eds.) *JALT2005 Conference Proceedings*. Tokyo: JALT.

We all know the metaphor of the vine and the branches. The vine is the essential or most important part and the branches are the less important part. In this paper, I would like to use this metaphor to describe the situation we face in language testing. I would like to ask, what in testing is the vine and what are the branches.

I have come to consider the vine and branches metaphor out of my experience in helping to draft the International Language Testing Association (ILTA) Code of Ethics and both the Japan Language Testing Association (JLTA) and ILTA Code of Practice. I realized that there are many ethical issues that need to be included in a code of ethics and even more points that need to be included in a code of practice. But are all of these issues and points equal? Are there some issues that are more important than others? These are very important questions because we often get into the situation in which ethical concerns clash. This clash is not unique to language testing. The issue of the mother's health versus the right to life of the unborn fetus is a famous one in the abortion debate. Whether we like it or not, we must make decisions in cases like this.

Particularly, in language testing ethics, we must ask if there is an ethical concern that matches the words often falsely attributed to the Hippocratic oath, *above all, do no harm*. Many scholars believe that the phrase originated with Hippocrates but comes from another of his writings, Epidemics, Bk. I, Sect. XI. The Greek

there can be translated as: "As to diseases, make a habit of two things—to help, or at least to do no harm." [For an in-text citation please provide a page or paragraph number?] Although the words above all, do no harm are not used in the oath, they reflect the core concern. Other issues are mentioned in that code, but this idea is central. It is the vine. All the others are branches.

I would like to argue that there is such a central ethical concern in language testing as well and that it is roughly the same as the one in medical ethics. Although it is not stated in so many words in the ILTA Code of Ethics, I believe that the central ethical imperative in that code is *above all, do no harm to the test taker*. Test developers and test users (these administrating the test, scoring it, analyzing the results, and, especially those making decisions based on the test results) must always ask themselves if their actions do no harm to the test taker.

I would like to look at two cases and try to work through the ethical considerations in both to see if it is possible to apply the rule of do no harm to the test taker. The first case is a fairly typical one in Japan. A company needs to decide which employees to post abroad and decides, reasonably enough, that the results of a test of the language in which business must be conducted abroad (usually English) will be used to make the decision. Primarily because of cost considerations, the company has its employees take a language proficiency test totally comprised of multiple choice items and selects the person who gets a score which, the testing company claims, indicates a proficient user of the language.

The question I would like to pose is this. Can we construct an argument, based on sound measurement principles, which will show that, in the case I have just described, no harm would be done to the test taker? Or to put it in more realistic terms, can we figure out the degree of risk that harm might be done to the test taker?

Let me add one more piece of information. The test company reports a correlation of 0.7 between their multiple-choice test and a test of spoken production. It seems to me that, with this last piece of information, we are ready to compute the degree of risk that the test taker might be harmed. What the company is asking is if this employee has the ability to both understand and to speak the language needed overseas. Let us assume for the moment that the multiple-choice test gives a reasonable estimate of the employees' passive ability—particularly the ability to understand the spoken language. However, we have to use the correlation figure given by the testing company, (0.7), to estimate the employee's ability to speak the language. To estimate the performance on one test using the results of another you must square the coefficient of correlation. In this case we get a figure of 0.49. This means that if we want to estimate the person's speaking ability from the results of the multiple-choice test, we will be right only about one time out of two. That means that we have a 50% chance of getting a mistaken estimate of the person's spoken ability and thereby doing harm to that test taker.

Notice that, in creating this argument, I didn't rely on my feeling about the usefulness of multiple-choice tests or criticism of the design of the test or of the individual items in it. Remember that I assumed that this particular multiple-choice test was a good measure of passive language ability. This second point, the quality of the test, obviously must be

considered in creating such an ethical argument. But in this particular case, I could show that there was a high risk of doing harm to the test taker even if I assumed the test was a good one. However, I do not believe that the initial point I mentioned, my feelings about multiple-choice tests, should enter into the construction of ethical arguments. We need to work from sound measurement principles, not our subjective feelings.

The next case is similar to the first except that, instead of relying entirely on a multiple-choice test, the company uses the multiple-choice test to screen employees. Those performing well enough on the multiple-choice test (those the testing company says have sufficient passive ability to have a chance of passing a face-to-face test of speaking ability) are allowed to take a test of their ability to speak the language. The testing company can demonstrate that the rater reliability of the test of speaking is above 0.9. Reliability is a measure of the consistency of the test results. In this case, the degree of agreement between the marks given by the two raters to the same test taker performance. The company also conducted a validity study in which a sample of prior test takers who received a rating of *able to communicate in the work situation abroad* and were subsequently sent abroad were studied to see that they were in fact able to perform as expected. Validity asks if the test is measuring what it is supposed to measure. In this particular case, the validity of the test was checked by comparing the predictions made on the basis of the candidate's test performance with that person's actual performance on the job.

With this information in hand we are ready to look at the ethical argument that can be constructed in this case. In this second example, we do not have to estimate the test taker's ability using a test that measures some other ability, but there are at least three issues that must be considered in deciding the degree to which this test battery could do harm to the test taker. The first is the quality of the multiple-choice test, the first stage in the testing procedure. Recall that this test is the gatekeeper for the second face-to-face test of speaking. A gate-keeping device can fail in two ways. It can shut out people who should be allowed in and it can also fail by allowing in people who should not be let in. In the particular case we are considering, the first type of failure is the more serious. If the multiple-choice test wrongly denies a person who is capable of passing the face-to-face test the right to take that test, that person will miss the chance to be posted overseas and this could have negative consequences for that person's career. If the multiple-choice test mistakenly allows a person who does not have the ability to pass a face-to-face test to take such a test, that test taker must endure the pain of having to deal with test tasks that are beyond his or her competence. But this is clearly the lesser of the two evils.

The designers of this particular test battery considered the potential harm to the test taker of the two types of *mistakes* and decided to err on the side of making the second sort of error. It was decided to set the cut-off score low enough to reduce the risk of not allowing qualified test takers to go on to the second test—even if this meant making more of the second kind of error. The designers decided to run the risk of allowing some unqualified test takers to go on to the second test (and perhaps causing pain to these test takers) in order to make sure that all qualified test takers got the chance to take the second test. This was an ethical question and the consequences of setting the higher or lower cut-off had to be

weighed. In the case of this test battery, the potential dangers of making the two sorts of errors were considered and the decision I have just reported was reached.

So far we have discussed only one aspect of the quality of the multiple-choice test—the setting of the cut-off score to decide who can go on to the face-to-face test. A more important question is whether or not this first test can really predict success in the face-to-face test. The correlation between the two tests is 0.79. This provides some evidence that the two tests are measuring somewhat different abilities (the correlation is not close to 1) and that success on the first test is a strong indication of potential success on the face-to-face test (the results of the two tests are reasonably positively correlated). So, this information together with the earlier discussion of the cut-off point shows that we can have some confidence that the risk of harming the test takers because of the poor quality of the multiple-choice test is rather low.

The other two ways this test battery might harm test takers is 1, if the reliability of the face-to-face test is low, and 2, if the judgments made on the basis of the test results (the decision whether or not to post the employee abroad) are not valid. As I mentioned earlier, rater reliability for the spoken test was above 0.9. We can see that the danger of harming test takers because of the inconsistency of the raters is very slight. The validity study checked to see that those who were posted abroad on the basis of their test battery results could actually perform abroad in the way expected. Interviews with the supervisors of such employees and with the native English-speaking colleagues who worked most closely with them indicated that they were performing as predicted. It was clear from this validity investigation that the test takers who were part of the study were not harmed by the test battery. Quite the opposite, it gave them the opportunity to further their careers.

In constructing this argument supporting the assertion that there is very little risk of this test battery harming test takers, I have relied on measurement procedures and principles. However, as we saw when we discussed the setting of the cut-off score for the multiple-choice test, decisions must be made. In this particular case, the test designers had to decide on the relative harm that would be caused by one of the two types of errors that had to be considered. We cannot avoid making these decisions or tradeoffs. But, as test developers, we must make it clear what decisions we have had to make and what the competing consequences are.

I have discussed elsewhere the relationship between a language testing code of ethics and a code of practice (see Thrasher, 2004), but I have come to the conclusion that the connection resides in the construction of ethical arguments. It is too simple to say that the use of a test is ethical if good testing practices have been followed. However, if our code of ethics provides the core principle of *above all, do no harm to the test taker*, the points of the code of good language testing practice can be used to construct an ethical argument indicating the degree of risk of harming the test taker.

Randy Thrasher is the Dean of The College of Liberal Arts, Okinawa Christian University, the President of the Japan Language Testing Association, and Secretary of the International Language Testing Association. His research interests include language testing, Relevance Theory, language teacher education.

References

Code of Ethics for ILTA http://www.iltaonline.com/code.pdf

International Language Testing Association, Draft ILTA Code of Practice http://www.iltaonline.com/code.htm

JLTA Code of Good Testing Practice http://www.avis.ne.jp/~youichi/COP.html

Thrasher, R. (2004). The Role of a Code of Practice in Language Testing Ethics. *Language Assessment Quarterly 1*(2)