LANGUAGE
TESTING

Test fairness and Toulmin's argument structure

Language Testing
27(2) 183–189
© The Author(s) 2010
Reprints and permission: http://www.sagepub.co.uk/journalsPermission.nav
DOI: 10.1177/0265532209349468
http://ltj.sagepub.com
SAGE

Antony John Kunnan
California State University, Los Angeles, USA

Keywords
test context framework, test fairness, test fairness framework, Toulmin argument structure

In this response, I will focus on the fairness and Toulmin's argument structure, Xi's proposal, and the challenges the proposal brings.

Test fairness

Test fairness as a fundamental concept in the evaluation of tests has been in the forefront of discussions in the field of language assessment from the late 1990s. An early such plea was a tentative fairness research agenda that I presented at the Language Testing Research Colloquium, Tampere, in 1996. I wrote then: 'Although validation studies are said to generally take on the role of investigating the fairness of tests and testing practices, an examination of about 100 validation studies shows that the themes addressed by the researchers are not particularly concerned with fairness' (Kunnan, 1997, p. 85). My interest in this issue was further invigorated as the theme of the LTRC, Orlando, in 1997 was 'Fairness in language testing.' But many scholars criticized the concept of test fairness as over-reaching and ambiguous and that such investigations were already part of the validation research agenda.

At about this time, the 1999 *Standards* (AERA, APA, NCME, 1999) was published with a section entitled 'Fairness in testing' with chapters on testing and test use, rights and responsibilities of test takers, test takers with linguistic diversity, and test takers with disabilities. This publication renewed interest in test fairness but as Xi rightly points out, test fairness was now seen as an additional test quality test developers were asked to pay attention to rather than as an integral part of test development and testing practice.

The Test Fairness Framework

I then put forward an ethics-inspired rationale for a Test Fairness Framework (TFF; Kunnan, 2004) with a set of first principles and sub-principles combining both the

Corresponding author:
Antony John Kunnan, TESOL Program, King Hall C 2098, 5151 State University Dr., California State University, Los Angeles, CA 90032, USA.
E-mail: akunnan@calstatela.edu

utilitarian and deontological ethical systems. The TFF, like Rawls' (1971) view of justice as fairness, is framed for a well-ordered society in which there is social cooperation between citizens who are free and equal and the primary goal is social and political justice.

The principles and sub-principles for the TFF are as follows:

1. the Principle of Justice: a test ought to be fair to test takers (sub-principle 1: a test ought to have comparable construct validity of score interpretations and decisions; sub-principle 2: a test ought not to be biased in terms of construct-irrelevant matters); and
2. the Principle of Beneficence: a test ought to bring good to society (sub-principle 1: a test ought to promote good to society; sub-principle 2: a test ought not to inflict harm to society).

These principles and sub-principles could then be used, I argued, in test evaluation where the specific focus could be on areas such as validity of test score interpretations and decisions, absence of test bias, test access, test administration and test consequences. Further, test fairness is also about fair testing practice so that tests are beneficial and not harmful to society.

Philosophical underpinnings

In addition, I proposed alongside the TFF that test developers and test users design a test and testing practice that is fair to all test takers without reference to a particular set of test takers. In other words, I suggested that test developers need to take a hypothetical and non-historical original position of test takers in which all test takers were situated as equals instead of designing a test that is fair for specific groups. In using this concept then, test developers and test users need to work with a veil of ignorance regarding particular test taker differences such as their race, ethnicity, gender, talents, wealth, social position, social and philosophical views, and so on. This conceptualization is borrowed from Rawls' (1971, 2001) *original position*, which is the view that when considering justice as fairness what is required is a veil of ignorance regarding members of society so that just and fair practices can be devised (without consideration for any group).

Toulmin's model of argument structure

The Toulmin model of argument structure (1958/2003), one of the taxonomies of argument representation schemes, used by Xi, is a method of practical reasoning with a structured macrostructure of arguments. It has the following six categories: the *claim* or assertion or conclusion which is a statement including any *qualifiers* or certainty or limits, the *evidence* or grounds or data that support the claim, the *warrant(s)* or principle or authority or reasoning connecting the evidence to the claim, the *backing* or reasons or assurances or theory, if necessary for warrants, and the *rebuttal(s)* or exception(s) to the claims. A typical Toulmin argument structure is often laid out in a diagrammatic structure that captures the relationship among the argument categories with logical connectors (so, unless, since, etc.) that helps preserve the logical nature of the argument. In general, an argument structured in such a form is expected to unfold strengths, weaknesses and limits of claims as opposed to philosophy's prior value attached to the analysis of syllogisms based on premises and conclusions. This model has been influential in cognitive science, legal argumentation, educational measurement (Mislevy, 1996; Kane, 2006) and language assessment (Bachman, 2005). Kane (2006) describes the Toulmin model:

> Toulmin treated his model as a dialogue between an advocate for a claim and a challenger. The advocate makes a claim based on an inference. The challenger can question the warrant for the inference or the appropriateness of applying the warrant in a particular case. If the warrant itself is challenged, its backing can be brought forward. The nature of this backing will depend on the nature of the warrant and the nature of the objection to it. (p. 28)

Based on this description, the Toulmin model should work well with a claimant and a challenger (as in a court) but when a research agenda or test evaluation is being planned or executed, the researcher's mind will have to imagine the different challenges making it difficult to identify and assemble the warrants, backing, qualifiers and rebuttals.

Test fairness research and applying Toulmin's argument structure

Xi proposes an approach to investigating test fairness to guide practitioners and illustrates it with the TOEFL iBT. Xi re-presents the six types of categories or inferences as modified by Chapelle et al. (2008) and, Bachman (2005) and Kane's (2006) argument-based approach. As a detailed review of Xi's proposal is beyond the scope of this commentary, a few key points will be discussed.

Test fairness research

Xi characterizes test fairness in terms of three prevailing views with examples: (1) fairness as a relatively independent test quality or general testing practice that is not clearly connected to validity (example, 1999 *Standards*); (2) fairness as an overarching test quality that consists of different facets including validity (example, Kunnan, 2004, 2008), and (3) validity as the fundamental test quality that links fairness directly to it (examples, 1999 *Standards*, Willingham & Cole, 1997). Xi dismisses the first two views: the first view 'does not provide a mechanism for prioritizing them and for weighing one piece of fairness evidence against another' (p. xxx) and the second view: 'current validation frameworks … have provided means to address all the fairness qualities proposed in Kunnan (2004) in a coherent way within the framework of a validity or assessment use argument. It does not seem necessary to treat them as separate facets of fairness' (p. xxx). These assertions help Xi to take up the third view as appropriate for her approach and illustration. Xi hitches this 'new' approach to an argument-based approach; thus coming up with *a fairness argument in a validity argument.*

In general, I have doubts as to the general usefulness of this approach as applied to language assessment. First, I do not see how current validation frameworks provide the means to investigate all areas of test fairness. Xi's view of test fairness is similar to the criticisms made in the early 1990s when the concept of test fairness was first proposed – that such a concept was under the jurisdiction of validity. Second, if current validation frameworks can cover fairness matters too, why would a fairness argument need to be positioned within a validation argument? Third, nesting or embedding fairness within validity of test score interpretations and decisions is a disservice to both concepts as the focus of research agendas could be diffused and confused. Fourth, reducing fairness matters into a series of rebuttals of validation studies diminishes the role given to fairness, as it can only react to validation arguments rather than set its own agenda.

In addition, in terms of priorities in my test fairness framework, I was looking forward to reading how Xi would set priorities for fairness investigations. Xi states 'we need to anticipate what the potential weaknesses are in the TOEFL iBT test fairness argument' (p. xxx) and 'focusing on the weakest areas in the argument would help maximize the use of resources, as it is typically not possible to address all the potential fairness issues' (p. xxx). But these statements do not lay out the priorities for research studies either indicating that priorities are best identified within a specific test context.

Thus, while current validation frameworks can adequately address narrower scope matters such as domain or content representativeness and comparable validity for all groups (argued by Willingham and Cole, 1997 and by Xi), it will be harder to address issues of fair testing practice.

Applying Toulmin's argument structure

As mentioned earlier, Toulmin's argument structure provides a logical approach for organizing claims, evidence, warrants, backing, qualifiers, and rebuttals in a test evaluation exercise.

But I see four challenges to the feasibility of using the Toulmin model. First, while claims, warrants, backing, qualifiers, and rebuttals seem relatively straightforward and appealing in the demonstration examples with singular facts (of a person's citizenship claims) or simple categorical statements in Toulmin's examples, there are some difficulties when we have competing claims. As an illustration, here are a few claims that have been made regarding the US Naturalization Test, depending on whether the claim is made by government officials, conservative or liberal political commentators, testing professionals, citizenship course instructors, or citizenship applicants: (1) the test assesses English language ability and US history and government; (2) the test assesses facts and figures about US history and government; (4) the test assesses civic integration and nationalism; (5) the test assess whether applicants will be productive and useful Americans; (6) the test assesses patriotism to the USA; (7) the test does not assess anything; it is just a hurdle to keep out people the government does not want. Once a decision is made as to which of these claims is going to be tested (or whether all of them are going to be tested) in an argument structure, then relevant evidence, warrants, and backing need to be identified and assembled but without the claimants' challenges of the warrants and backing.

Second, does the model easily allow for an integrated argument? Specifically, is it possible the six separate categories might carve up the main argument into separate categories thus failing to provide the synthesis needed to defend a claim? For example, is it possible that when evidence from several different studies (exploratory factor analysis, structural modeling, conversational analysis, ratings, etc.), are plugged in as evidence, warrants, backing or rebuttals in the argument structure, they could become separate arguments rather than one synthesized argument?

Third, could both explicit and implicit or stated and unstated claims be equally testable or would there be too much classificatory ambiguity? For example, in the US Naturalization test, the explicit claim is that it is a test of English language ability and US history and government. What evidence is there to support this claim? But, the implicit claims behind the test are 'civic integration' and 'civic nationalism' both of which are vague at best. What is the evidence, warrant or backing for these claims? Or, is the test merely satisfying statutory requirements? If this is the case, how would one go about testing this claim using the Toulmin model?

Finally, could the argument emerging from the argument structure be understandable to general researchers and readers? And, would it be suitable for all disciplines and for all research purposes? For example, how would Toulmin model deal with claims such as 'Latin jazz is the best music in the world' or 'The Taj Mahal is the most aesthetically pleasing building in the world', or 'the IELTS is better than the TOEFL' where the claims, evidence, warrants and backing are not singular facts or categorical statements but could be statements of relative belief?

To sum up the challenges, I wonder what has been gained from using the Toulmin argument structure. It does not seem to have broken the shackles of the older "premises and conclusion" model of argumentation.

Beyond fairness and argument structure

Realizing that the TFF framework is a micro-analytic approach limited in scope to evaluating a test (and not the contexts surrounding it), I put forward a Test Context Framework (Kunnan, 2005, 2008), a macro-analytic approach, that would consider the wider context of tests (political, economic, educational, social, cultural, technological, and legal). In this context, how would an argument-based structure be useful in such an analysis? In a recent examination of the motivations and impacts of citizenship testing in the US in the 20th century, I examined the political and social consequences of the testing practice (Kunnan 2009a, 2009b). I used a non-formal analysis and I argue that I would not have been served well with the Toulmin model as I had to make many connections and interpretive judgments that by structure the Toulmin model may have had difficulty accommodating.

But returning to my view that a test and testing practice ought to be fair and beneficial to society gives me the sense that the main issue here is inherently an ethical one.

Thus, the argument or defense of a test and testing practice has to be made on ethical grounds. Toulmin (1972) in *Human Understanding* has this to say about how to approach ethical questions:

> In any culture and generation, men acknowledge the authority of a dozen inherited approaches to ethical questions. Each of these approaches has its own rubric – 'as a matter of self respect/ morality/loyalty/etiquette /integrity/equity/religious commitment/simple humanity…' – and each defines a particular set of issues, considerations, and modes of argument. In any chosen culture and generation, furthermore, men do not merely continue applying all these different considerations and arguments in exactly the same way as their forefathers; they also attempt to refine their application, and to reorder their relative priorities, in light of the changing needs and conditions of life. (Cited in Stygall, pp. 10–11)

Thus, a decontexualized argument structure like the Toulmin argument structure may not be necessary or even useful in all test evaluation cases. In fact, research based on using argument structures may do well to include non-formal analysis (like thinking outside the box; here, the Toulmin box). In conclusion, Xi has provided an interesting proposal that is both thought-provoking and challenging to address. I hope in the future we will continue to see discussions on test fairness and evaluations of argument structures (see Newman & Marshall, 1992, for an example) as we try to understand what test fairness means and how best to conduct test fairness research.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: Author.

Bachman, L. F. (2005). Building a test use argument. *Language Assessment Quarterly, 2,* 1–34.

Chapelle, C. A., Enright, M. K. & Jamieson, J. M. (2008). (Eds.), *Building a validity argument for the Test of English as a Foreign Language.* Mahwah, NJ: Lawrence Erlbaum**.**

Kane, M. T. (2006). Validation. In Brennan, R. L. (Ed.), *Educational measurement,* 4th ed. (pp. 18–64). Washington, DC: American Council on Education/Praeger.

Kunnan, A. J. (1997). Connecting validation and fairness in language testing. In A. Huhta et al. (Ed.), C*urrent developments and alternatives in language assessment* (pp. 85–105). Jyväskylä, Finland: University of Jyväskylä.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context* (pp. 27–48). Cambridge, UK: Cambridge University Press.

Kunnan, A. J. (2005). Language assessment from a wider context. In E. Hinkel (Ed.), *Handbook of research in second language learning* (pp. 779–794). Mahwah, NJ: Lawrence Erlbaum.

Kunnan, A. J. (2008). Towards a model of test evaluation: Using the Test Fairness and Wider Context frameworks. In L. Taylor & C. Weir (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity* (pp. 229–251). Cambridge, UK: Cambridge University Press.

Kunnan, A. J. (2009a). Testing for citizenship: The U.S. Naturalization Test. *Language Assessment Quarterly, 6*, 89–97.

Kunnan, A. J. (2009b). Politics and legislation in citizenship testing in the U.S. *Annual Review of Applied Linguistics, 23*, 37–48.

Mislevy, R. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*, 379–416.

Newman, S. & C. Marshall (1992). Pushing Toulmin too far: Learning from an argument representation scheme. Palo Alto, CA: Xerox Palo Alto Research Center.

Rawls, J. (1971). *A theory of justice.* Cambridge, MA: Harvard University Press.

Rawls, J. (2001). (Editor: E. Kelly). *Justice as fairness: A restatement.* Cambridge, MA: Harvard University Press.

Stygall, G. (n.d.) *Toulmin and the ethics of argument fields*. Washington, DC: Office of Educational Research and Improvement.

Toulmin, S. (1958/2003). *The uses of argument.* Cambridge, UK: Cambridge University Press.

Toulmin, S. (1972). *Human understanding.* Princeton, NJ: Princeton University Press.

Willingham, W. W. & N. Cole (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum.

Copyright of Language Testing is the property of Sage Publications, Ltd. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.