JLTA Journal, vol. 23: pp. 3–13, 2020

DOI: 10.20622/jltajournal.23.0_3
Print ISSN 2189-5341 Online ISSN 2189-9746

Evaluating Language Assessments From an Ethics Perspective: Suggestions for a New Agenda

Antony John KUNNAN
Duolingo, Inc.

Introduction

The dominant 20th century approach to the evaluation of language assessments was the *Standards*-based approach. The *Standards* most evaluators referred to are the American Psychological Association (APA), American Educational Research Association (AERA), National Council on Measurement in Education (NCME) *Standards* (1999, 2014). These standards (mainly a list of test qualities such as validity and reliability, and of late, consequences and fairness) were developed from best practices at assessment institutions and had loose connections to theories of educational and psychological measurement. The "Test Usefulness" concept proposed by Bachman and Palmer (1996) was a popular example of the *Standards* approach. In the early part of the 21st century, Kane (1992) and Bachman and Palmer (2010) proposed an Argument-based approach using Toulmin's way of structuring arguments with claims, warrants, backing and rebuttals. This approach provided a framework for evaluating language assessments. Bachman and Palmer's (2010) "Assessment Use Argument" (AUA) is an example of this approach.

While both approaches provide ways for researchers to conduct evaluations, they have a weakness, and that is they generally lack an articulated philosophical grounding. This lack of philosophical grounding can be seen in the *Standards* approach in which why the listed standards are important and not others is not articulated. In the *Argument* approach, what aspects are to be included as claims and warrants is left the assessment developer with the evaluator following them which is a critical problem. To remedy this situation, I am proposing an *Ethics*-based approach to assessment evaluation. The framework that implements the approach harnesses the dual concepts of fair assessments and just institutions leading to the *Principle of Fairness* and *Principle of Justice,* respectively.

An Ethics-Based Approach

An Ethics-based approach draws on the perspective from the world of moral philosophy, in which an ethic or ethical knowledge can be used to morally justify individual and institutions practices. This support can empower both approaches in helping with general questions related to assessment and assessment practice: For example, these moral questions can be asked:

(1) Does every test taker have the right to a fair assessment? Is this rule inviolable? Are rights of test takers to a fair assessment universal or only applicable in states that

provide equal rights? Is it adequate that most test takers are assessed fairly while a few are not?

(2) What responsibilities does a test developer or test score user have? Would it be appropriate to use a cost–benefit analysis to evaluate whether assessments should be improved or not? If harm is done to test takers, does such harm need to be compensated?

(3) Would the rights of test takers to a fair assessment be supported in authoritarian states that do not provide for equal rights? Would institutions in such states feel less compelled to provide a fair assessment?

(4) Should assessment developers and users be required to offer public justification or reasoning? Should they present their justifications for assessments backed by research findings in appropriate forums? Should an assessment be beneficial to the society in which it is used?

(5) Should assessment institutions be just in their approach?

Secular philosophers from centuries ago including Socrates, Plato and Aristotle have searched for the meaning of justice. The main proponents, however, who addressed these matters were Enlightenment philosophers such as Locke, Hume, Bentham, Mill and Sidgwick. These philosophers were called utilitarianism and their general moral theory holds that the rightness or wrongness of an action is determined by the balance of good over evil that is produced by that action. Thus, rightness of actions (by individual and institutional) should be judged by their consequences (caused by the actions). This important aspect of utilitarianism is termed consequentialist thinking in which outcomes of an event are used as tools to evaluate an institution. Another doctrine of utilitarianism is the Greatest Happiness Principle; it promotes the notion that the highest principle of morality is the greatest happiness for the greatest number of people: to maximize utility and to balance pleasure over pain. As a result, the utility principle would trump individual rights.

Implementing utilitarianism in the field of assessment could mean that decisions about an assessment may be made solely on utility and consequences. For example, if an assessment brought in a great deal of revenue because of large numbers of test takers taking an assessment, the assessment could be considered successful. In addition, if the consequences of the assessment were positive for a large majority, then the assessment could be considered beneficial to the community. However, maximizing happiness or minimizing unhappiness can result in sacrificing fairness and justice. For example, suppose an assessment was biased against a group of test takers, and to improve the current version or to develop a new assessment would entail a great deal of expenditure. This expense, if carried out, then would result in everyone paying more for the assessment and causing harm to all. One forced choice could be that the assessment be continued the way it is without any improvement. Strict utilitarians, in this case, would argue that these are bad choices and the lesser harmful of the two options would need to be chosen. Such utilitarians would hold that even if an

assessment is biased against a group and fairness and justice may have to be sacrificed, we will have to just live with the assessment without any improvement. It would maximize happiness and minimizes unhappiness.

Another way of thinking of ethics emerged with deontological (duty-based) ethics pioneered by the works of Immanuel Kant (1724–1804). Kant argued that to act in the morally right way, people must act from duty, and unlike utilitarianism, it was not the consequences of actions that made actions right or wrong but the motives of the person who carried out the action. His assertion was that there is a single moral obligation called the Categorical Imperative derived from duty and that people are naturally endowed with the ability and obligation toward right reason and acting. The Categorical Imperative can be considered an unconditional obligation.

In addition, William Ross (1877–1971) offered seven prima facie duties that need to be considered when deciding which duty should be acted upon. Three of them relevant for this discussion were: Duty of beneficence (to help other people to increase their pleasure, improve their character, etc.), the Duty of non-maleficence (to avoid harming other people), and the Duty of justice (to ensure people get what they deserve).

Rawls (1971), in "A Theory of Justice" formulated a theory and principles of fairness and justice in which he argued that fairness is foundational and central to justice and therefore it is prior to justice. To quote from Sen's (2009) summary of Rawls's work: In the Rawlsian theory of "justice as fairness", *the idea of fairness relates to persons* (how to be fair between them) whereas the Rawlsian principles of justice are applied to *institutions* (how to identify "just institutions," p. 72).

Principles of Fairness and Principles of Justice

Adopting many ideas from these philosophers, I proposed two Principles of Fairness and Principles of Justice:

Principle 1 — The Principle of Fairness: An assessment ought to be fair to all test takers, that is, there is a presumption of treating every test taker with equal respect.

Principle 2 — The Principle of Justice: An assessment institution ought to be just, bring about benefits in society, promote positive values, and advance justice through public justification and reasoning.

These principles, based mainly on deontological thinking, could guide evaluation of language assessment professionals to include the concepts and applications of fairness and justice. This focus also alters the dominant view of examining assessments and assessment practice (as seen through the focus on validation and reliability studies) to how assessments relate to test takers and their community (beneficial consequences).

Early applications of an ethics-based approach include a few general ideas that were proposed by educational and language assessment researchers. Davies' (1977) "Test virtues"

approach made an early argument for "test virtues," reflecting Aristotelian virtue ethics. Davies argued that test developers or agencies ought to act as moral individuals or agents who have ethical principles by which they operate and therefore do the right thing for the right reasons.

Kunnan (1997) proposed a fairness agenda after reviewing 100 validation studies conducted by well-known language assessment researchers. He argued that:

> a social postmodernist view, in contrast, would value validation research that is attentive to social and cultural difference, not just by learning about differences among test takers or indulging in an easy relativism which in practice might result in not taking difference seriously, but by engaging in a research program that would incorporate social and cultural difference within the validation process (p. 93).

Willingham and Cole (1997), in their study of gender and fair assessment, argued that "test fairness is an important aspect of validity…anything that reduces fairness also reduces validity…test fairness is best conceived of as comparability in assessment; more specifically, comparable validity for all individuals and groups" (pp. 6–7). Using the notion of comparable validity as the central principle, Willingham and Cole (1997) suggested three criteria for evaluating the fairness of a test: "comparability of opportunity for examinees to demonstrate relevant proficiency, comparable assessment exercises (tasks) and scores, and comparable treatment of examinees in test interpretation and use" (p. 11).

FairTest, a non-profit organization in the US, has devoted its entire work to the cause of fair assessments. Its mission is to advance "quality education and equal opportunity by promoting fair, open, valid and educationally beneficial evaluations of students, teachers and schools." FairTest also works to end the misuses and flaws of testing practices that impede those goals.

A New Research Agenda

A research agenda articulated in terms of the Principle of Fairness and the Principle of Justice is outlined below. The Principle of Fairness refers to fairness of assessments in terms of test takers and the Principle of Justice refers to just institutions that are responsible for administering assessments and making decisions based on test takers' scores. Specifically, ethical principles lead to general claims and these lead to sub-claims. Figure 1 shows the flow of the principles, claims, warrants, etc. In the example below, these claims and sub-claims are written with a fictitious *Assessment X* in mind.

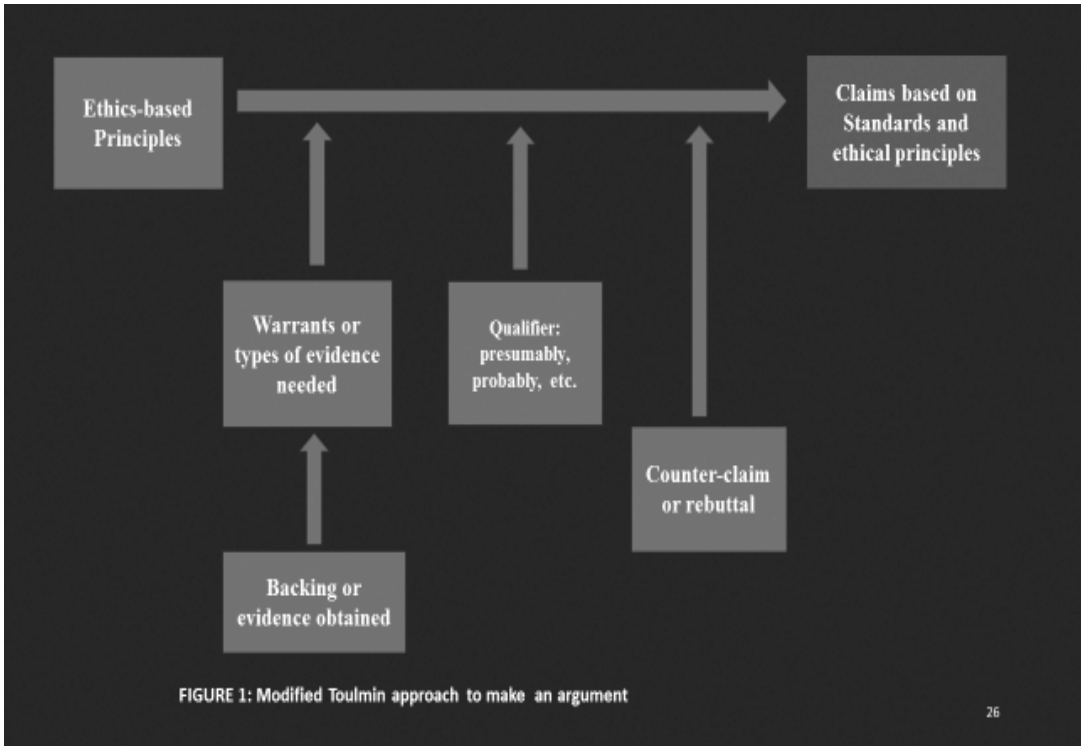

Figure 1. Modified Toulmin approach to evaluate an assessment.

Principle 1

The Principle of Fairness: Assessment (named X) ought to be fair to all test takers, that is, there is a presumption of providing every test taker with equal opportunities to demonstrate their abilities.

General Claim: *Assessment X* is fair to all test takers.

Sub-claim 1: Prior to taking *Assessment X*, adequate opportunity to learn and prepare is provided.

Sub-claim 1.1: Prior to taking Assessment X, adequate opportunity to learn is provided.

Sub-claim 1.2: Prior to taking Assessment X, adequate time for preparation is provided.

Sub-claim 1.3: Prior to taking Assessment X, adequate familiarity and practice with new technology is provided.

Sub-claim 1.4: Prior to taking Assessment X, adequate semiotic embodied experience in the domain of the assessment is provided.

Sub-claim 2: *Assessment X* is both meaningful and consistent.

Sub-claim 2.1: Assessment X is meaningful as it is representative of the blueprint and specifications, or curriculum objectives.

Sub-claim 2.2: Assessment X is meaningful as it represents the constructs of interest in the assessment.

Sub-claim 2.3: Assessment X is meaningful as it represents the language variety, content, and topics of interest in the assessment.

Sub-claim 2.4: Assessment X is meaningful in that it can predict performance in terms of external criteria.

Sub-claim 2.5. Assessment X is consistent within sets of items/tasks in terms of different constructs.

Sub-claim 2.6: Assessment X is consistent across multiple assessment tasks, forms and/or occasions of assessments (in different regions, offices, and rooms).

Sub-claim 2.7: Assessment X is consistent across multiple examiners involved in the assessment.

Sub-claim 3: *Assessment X* is free of bias.

Sub-claim 3.1: Assessment X is free of bias in terms of content, topic or language variety across test taker groups.

Sub-claim 3.2: Assessment X is free of differential performance by different test taker groups of similar ability (examples, in terms of gender, age, race/ethnicity, and native language).

Sub-claim 3.3: Assessment X score-interpretation is based on defensible standard-setting procedures.

Sub-claim 4: *Assessment X* provides appropriate access and administration.

Sub-claim 4.1: Assessment X is affordable to test takers.

Sub-claim 4.2: Assessment X is administered at locations that are accessible to test takers.

Sub-claim 4.3: Assessment X is accessible to test takers with disabilities and has appropriate accommodations.

Sub-claim 4.4: Assessment X is accessible to test takers whose L1 is not English for subject matter tests (examples, science, mathematics, computer science, etc.).

Sub-claim 4.5: Assessment X is administered uniformly to test takers.

Sub-claim 4.6: Assessment X is administered with appropriate security that limits fraud or security lapses.

Sub-claim 4.7: Assessment X provides scores for decision-making that are defensible.

Principle 2

The Principle of Justice: The assessment institution that administers Assessment X ought to be a just institution as the assessment ought to be beneficial to society.

General claim: *Assessment X* is administered by a just institution.

Sub-claim 1: Assessment X is beneficial to the immediate community and larger society.

- Sub-claim 1.1: Assessment X makes decisions that are beneficial to immediate stakeholders (e.g., test takers, instructors in citizenship courses in community colleges, and college administrators).
- Sub-claim 1.2: Assessment X makes decisions that are beneficial to test-takers (in terms of gender, age, race/ethnicity, size of community).
- Sub-claim 1.3. Assessment X makes decisions that are beneficial to the instructional program (e.g., teaching-learning, and the learning environment, also known as "washback").
- Sub-claim 1.4. Assessment X makes decision that are beneficial to the wider stakeholders (e.g., school district, community, province/state, country).

Sub-claim 2: *Assessment X* provides positive values and advances justice.

- Sub-claim 2.1: Assessment X has provision for administrative remedies to challenge decisions such as rescoring or re-evaluation.
- Sub-claim 2.2: Assessment X has provision for legal challenges related to decision-making.
- Sub-claim 2.3: Assessment X makes decisions that are not detrimental to test taking groups and corrects existing injustice (if any) to test taking groups.
- Sub-claim 2.4. Assessment X institution promotes positive values and advances justice by providing public justification and reasoning for the assessment.

The claims and sub-claims that are the direct result of an ethics-based approach include the following: Opportunity-to-Learn and Free of bias under the Principle of Fairness and beneficial assessment and positive values and advancement of justice under the Principle of Justice. These claims are typically not pursued under the Standards-based and Argument-based approaches except for Bachman and Palmer's focus on equitable decisions and beneficial consequences. At first glance, these claims and sub-claims may also seem to be obvious to researchers who use standards-based and argument-based approaches. But these unique principles are operationalized into claims and sub-claims that can be evaluated in terms of collected evidence. Further, they are articulated in a framework with supporting philosophical positions. These sub-claims, if they are part of assessments, can be examined through traditional quantitative and qualitative research methods.

Implementing the New Research Agenda

A keyway for any organization to contemplate the application of these principles is to develop a plan for the new research agenda. This can best be done by reflecting on the teaching curriculum that is used in training teacher-researchers and professional researchers. There are three general approaches in which the focus and purpose are different. Here is an outline of the approaches.

The Traditional Engineering-Oriented Curriculum

This approach includes:

(1) An overview of skills and components to be assessed (listening, speaking, reading writing; pronunciation, grammar, vocabulary; pragmatics, integrated skills, etc.)
(2) Techniques for test development, creation, and scoring (item writing and revision, using different response formats, writing rubrics for scoring, designing score reporting formats, etc.)
(3) Skills for improving, editing, and assembling of items and tests (planning revision cycles, writing protocols for item banking, etc.)
(4) Research themes: reliability (internal consistency, inter-rater), content, construct and predictive validity, and item and test bias (differential item/test functioning analysis)
(5) Research methods: quantitative (classical true score theory, item response theory) and qualitative (introspective analysis, conversational analysis) approaches for research (procedures for item and test difficulty, discrimination, bias, analyzing cognitive and affective processes in test taking, etc.)
(6) Training with software: SPSS, ITEMAN, and SAS for descriptive and inferential statistics; FACETS and WINSTEPS for analyzing ratings; MULTILOG for bias analysis, etc.
(7) Technical and score user manuals (writing parts of a technical manual that includes a description of the test; administrative, scoring, and reporting details; accommodations for test takers with disabilities; a report of research studies that support the claims of the test; samples of tests, scoring rubrics, and score reports; and pricing of different services—standard prices, rush score reporting, human scoring, detailed score report, and diagnostic feedback, etc.).

The Innovative-Design Curriculum

This approach includes:

(1) Innovative design in items and tests and scoring and reporting (integrated tasks such as listening–writing, reading–speaking, tasks that match professional work such as assessing the English language ability of aviation professionals, etc.)

(2) Use of new technologies (tasks that involve audio-video, multi-media, multi-modal, automated scoring of writing and speaking, tests on tablets, etc.)
(3) New needs and uses (immigration, citizenship, asylum, forensic purposes, etc.)
(4) Research themes: authenticity, instructiveness, cognitive diagnostic feedback, dynamic assessment, etc.
(5) Research methods: new analyses using structural equation modeling, hierarchical linear modeling, multi-level modeling, etc.
(6) Training with advanced software: AMOS, EQS, Mplus for modeling, etc.

The Ethical-Critique Curriculum

This approach includes:

(1) Understanding of the history of language assessment (from the Chinese Civil Service Examinations, Le Baccalaureate, and Abitur to popular modern language assessments)
(2) Understanding the political motivations and legal bases for assessments (the U.S. Naturalization Test, similar tests in the U.K., Australia, Germany, South Korea, etc.)
(3) Understanding the social and cultural assumptions of assessments (knowing the semiotic domain and having embodied experiences)
(4) Understanding the philosophical underpinnings of assessments (utilitarian, social contract, deontological, pragmatist, and humanist approaches)
(5) Working with hypothetical scenarios or case studies of assessments that need judgments using moral or ethical philosophy
(6) Using ethical principles to evaluate assessments (parochial versus global ethics, etc.)
(7) Writing defences and critiques of assessments and assessment practice (assessment reviews)

It may be obvious that the focus of most training programs (certificate, BA, MA, or PhD) would be on the first approach, although a few programs would address the second approach. And, unfortunately, aspects of the third approach are ignored or not offered regularly. This could be due to several reasons: The first approach lays the foundation of training and therefore has to be included in all programs. The focus of programs could be based on the expertise of the faculty of the program (which is generally in the first strand). The second approach has to do with innovation in design, tasks, and research and is based on new purposes, contexts, and technologies. The third approach is the newest and requires the most time to develop, as it is interdisciplinary with subjects in humanities and social sciences.

Fulcher (2012) offers an expanded working definition of language assessment literacy, which has been in the forefront of discussions on the topic: The knowledge, skills and abilities required to design, develop, maintain or evaluate, large-scale standardized and/or classroom-based tests, familiarity with test processes, and awareness of principles and

concepts that guide and underpin practice, including ethics and codes of practice. The ability to place knowledge, skills, processes, principles and concepts within wider historical, social, political and philosophical frameworks in order [to] understand why practices have arisen as they have, and to evaluate the role and impact of testing on society, institutions, and individuals. He shows the different layers of knowledge working together: knowledge of skills and abilities (termed practices); processes, principles, and concepts (termed principles); and historical, social, political, and philosophical frameworks (termed contexts). Although this framework does not completely overlap with the curricula presented earlier, the traditional engineering curriculum can be matched with the practices and the ethical critique curriculum can be matched with the contexts.

The challenge for language assessment training programs to offer a comprehensive approach incorporating the most important components of the three curricula of Fulcher's layers based on needs and orientation of the programs. It is hoped, however, that all programs will include elements of the ethical-critique curricula and/or a historical, social, political, and philosophical context so that students and professionals are aware of the context of their work, as well as their responsibilities (for example, like health care professionals such as medical doctors, pharmacists, and hospital nurses and engineers and computer scientists are increasingly asked to do). A training program based on this approach is likely to offer its trainees the necessary capabilities to develop a new research agenda that can encompass the two principles of fairness and justice.

Conclusion

The central thesis of this article is that an ethics-based approach to evaluation of language assessments has the scope to promote fairness of assessments and just institutions that are not clearly brought to the forefront in the Standards-based and Argument-based approaches. While the Standards-based approach provide the best educational assessment practice approach and the argument-based approach offers a useful and clear framework for evaluating claims, they do not harness an ethical foundation to motivate assessment developers and researchers. An ethics-based approach as the one proposed here can be a useful extension to the on-going debate of the *what* and the *why* principles that should be employed for language assessment development and evaluation. To achieve the goals of this new research agenda, a teaching curriculum that encompasses key elements that are critical need to be in place.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. The author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and*

psychological testing. The author.
Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.
Davies, A. (Guest Ed.). (1997). Introduction: The limits of ethics in language testing. *Language Testing*, *14*(3), 235–241. https://doi.org/10.1177/026553229701400301
Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, *9*(2), 113–132. https://doi.org/10.1080/15434303.2011.642041
Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527
Kunnan, A. J. (1997). Connecting fairness and validation. In A. Huhta (Ed.), *Current developments and alternatives in language assessment* (pp. 85–105). University of Jyväskyla.
Kunnan, A. J. (2018). *Evaluating language assessments.* Routledge.
Rawls, J. (1971). *A theory of justice*. Harvard University Press.
Sen, A. (2009). *The idea of justice*. Harvard University Press.
Willingham, W. W., & Cole, N. (1997). *Gender and fair assessment*. Lawrence Erlbaum Associates.