

A Critical Book Review of "Ethics and Context in Second Language Testing: Rethinking Validity in Theory and Practice" Edited by M. Rafael Salaberry, Albert Weideman and Wei-Li Hsu (2023)

Ali Momeni^{1*}

| ARTICLE INFO | ABSTRACT |
|---|---|
| <p>Article History: Received: July 2024 Accepted: August 2024</p> | <p>The essential requirement of validity in second language assessment and testing is of paramount importance since it guarantees that tests properly gauge the targeted constructs and produce propitious interpretations. In this light, there seems to be a pressing need to carefully assess outdated conceptions of validity in English language testing to ensure they are in lockstep with current linguistic and sociocultural circumstances. This review examines the book "Ethics and Context in Second Language Testing: Rethinking Validity in Theory and Practice," edited by M. Rafael Salaberry, Albert Weideman, and Wei-Li Hsu (2023), and it unravels new horizons for expanding the comprehension of validity beyond traditional frameworks. The review underscores the primary subjects addressed in the book's content such as ethical issues, contextual factors affecting test design, and novel approaches for assessing validity in real-world settings. The review also stresses the book's critical analysis of existing paradigms and recommends a more nuanced perspective on validity that incorporates ethical considerations and contextual significance. To this end, the current review constitutes an important reference for researchers and practitioners to encourage them to critically examine and reframe validity in second language testing.</p> |
| <p>KEYWORDS Ethics in language testing Validation process Validity Validity in practice Validity in theory</p> | |

This informative and groundbreaking book edited by two well-regarded scholars in the field (e.g., M. Rafael Salaberry and Albert Weideman) provides an introduction to the theory and research of critical orientations toward language assessment, with an emphasis on socially contextualized and ethical considerations that arise in the backdrop of language testing and validation. The book propitiously discusses the inquiries about equality and social justice in teaching language in terms of empirical works, using both quantitative and qualitative methods. The editors provide a recap of crucial notions, theoretical concerns, and recommendations for investigations that advance the field. Language program managers, advanced learners, and scholars in applied linguistics, sociolinguistics, education, language policy, and related disciplines will find this book a valuable reference due to its novel outlook on language assessment. The book is organized into three parts, namely the ethical contextualization of validity, agency, and empowerment prompted by test adequacy and socio-interactive perspectives and assessment, with each part comprising a minimum of two to four chapters.

The first chapter, *Context, Construct and Ethics* by Rafael Salaberry and Albert Weideman, denotes a reconceptualization of the construct of language in terms of interactional, embodied, and multilingual turn in the history of second language acquisition. In this way, the chapter puts forward a

¹ PhD Candidate of Applied Linguistics, Allameh Tabataba'i University, Tehran, Iran.
Email: ali.momenniii@gmail.com; momeni_ali@atu.ac.ir (corresponding author)

'narrow' and a 'wide' view of validity (Schildt et al., 2023),"with the first referring to the very early view of validity as the characteristic of a test to test what it has set out to measure, i.e., to fulfill its purpose, the second to a position that includes the social and ethical impact of a test" (p. 9). Moreover, it is contended that language testing has progressed beyond the belief that a test is fair as long as the potential of incorrectly categorizing candidates has decreased to an acceptable level. Instead, the principles of critical language testing, which involve engaging those affected by test designs and ensuring that individuals who speak multiple languages and belong to minority language groups are given special consideration, must be implemented to address the issue of fairness in language tests to some extent.

The second chapter, *Validity and Validation: An Alternative Perspective*, by Albert Weideman and Bart Deygers, aims to provide a clear conceptual understanding of validity by distinguishing between the subjective process of validation and the objective characteristics of a language test. To obtain a clear understanding of the subject, it is necessary to construct a theory in applied linguistics known as "*primary applied linguistic artefacts*," which encompasses three key components: language policies and plans, language assessments and tests, and language curricula and courses. This chapter has contended that the technical design characteristics of a measuring instrument are essential but not enough to ensure validity. Likewise, it is stated that argument-based validity theory does not refute the existence of that argument; on the flip side, it emphasizes the significance of design-based conclusions and encourages test creators to clearly define score interpretations and back them up with evidence (e.g., data). Although it is acknowledged that argument-based validity theory promotes the importance of robust measurement for valid score usage (Kane et al., 2017), it should be emphasized that this theory may have led to the abandonment of some essential aspects of robust measurement in the design process. Moreover, the chapter unravels the reputation of a test with a critical concluding remark that "*in some sense, validity does reside in test instruments*", and it is therefore "*not just a trick of semantics . . . to say that one test is more valid than the other for a particular purpose*" (Davies & Elder, 2005, p. 798). In this respect, validity may well resurface as 'adequacy', 'effectiveness', or 'quality' which is evident in Messick's claim (1981, p. 18) that "*test validity is . . . an overall evaluative judgment of the adequacy and appropriateness of inferences drawn from test scores.*" The chapter also positions a more encompassing orientation toward validity that considers both objective design-based conclusions and the subjective subtleties of validation.

The third chapter, *The Racializing Power of Language Assessments* by Casey Richardson, serves as a potent criticism of language assessments, revealing their role in upholding white dominance and racial hierarchies. The author contends that standardized assessments perpetuate raciolinguistic ideologies by favoring some language types as "*legitimate*" while devaluing others, drawing on Messick's research on the social consequences of tests and Shohamy's idea of critical language testing. The chapter utilizes critical race theory to illustrate how language tests function based on the premise of interest convergence. This means that changes in language testing only progress when they coincide with the pursuits of white individuals or what is called "*white hegemony*." The chapter questions the prevalent conception that testing is fair and impartial. It discloses that the assumption of objectivity enables testing authorities to escape accountability for the negative impacts of their practices. The author advocates for significant modifications and changes, such as incorporating community involvement in assessments and recognizing students' various language needs and considerations. This approach strives to remove the bias in language assessments and live up to a more equitable assessment system. By using these critical recommendations, the chapter illustrates a compelling vision for reforming how assessments are performed, ensuring they are more inclusive and representative of all students' linguistic backgrounds.

The fourth chapter, *It's Not Their English: Narratives Contesting the Validity of High-Stakes Test*, by Gordon Blaine West and Bala Thiruchelvam, commences the discussion by conceptualizing language testing as a "*product and agent of cultural, social, political, education, and ideological agendas that shape the lives of individual participants, teachers and learners*" (Shohamy, 2001, p. 131). The chapter stresses that the validity of assessments is frequently assessed in a hierarchical mode, with validity being assessed either by policymakers or test developers. In addition, by applying the principles of critical language assessment, the chapter analyzes the legitimacy of the Computer-

based English Test (CBET), a high-stakes English language assessment. The viewpoint of test-takers and instructors at a university in South Korea is explored through qualitative interviews to get insight into the process of constructing and challenging validity from a bottom-up procedure. The study successfully emphasizes the perspectives of those most affected by the CBET, demonstrating how the validity of the exam is established based on the ethical stances and real-life encounters of test-takers and instructors. Nevertheless, the authors highlight the politically contentious aspect of their research, acknowledging that several instructors were disturbed by the apparent critique of the assessment.

The fifth chapter, *The Ethics of Potential of L2 Portfolio Assessment*, by Mitsuko Suzuki, rigorously analyses the ethical potential of L2 portfolio assessment from an outlook focused on power dynamics. The chapter argues that portfolios might be positioned as an alternative to typical standardized testing, with the aim of fostering a more ethical learning environment. On the other hand, the analysis of empirical research indicates that portfolios do not inherently empower second-language learners. The data demonstrate that students had varied responses to different portfolio activities. While several students appreciated the chances for introspection and constructive criticism from teachers, others perceived portfolios as being forced upon them, leading to heightened levels of anxiety. The learners' experiences were influenced by factors such as their competency level and institutional assessment needs. The chapter culminates with advocating a more discerning strategy for investigating and conducting portfolio assessments. It stresses the necessity of conducting negotiations between instructors and students and including other relevant parties to effectively utilize the empowering capabilities of this alternative assessment technique.

The sixth chapter, *Portfolio Assessment: Facilitating Language Learning in the Wild*, by Elisa Räsänen and Piibi-Kai Kivik, proposes an alternative method of assessing students in the classroom. This method involves using a portfolio that takes advantage of students' activities "beyond the classroom, in the wild". In this line, Eskildsen et al. (2019) point out that while the term "in the wild" is commonly perceived as the opposite of classroom learning, the phenomenon is more complex, and there exists a more "gradient" link between the two. It is stated that the implementation of portfolio assessment had a beneficial impact on students, as it motivated them to engage in real-life interactions as a means of learning actively and directed their attention toward specific learning opportunities within those interactions. The research revealed that the portfolio encouraged students to actively engage in existing relationships in the target language, form new connections using the target language, and utilize the portfolio as a tool for self-reflection on their learning. The reflections highlighted that students focused on language factors such as vocabulary and register, reflecting their growing metalinguistic awareness. The chapter concludes with a critical remark that:

language class assessment focusing solely on proficiency does not sufficiently address the objectives of cultural awareness and interactional competence; however, proficiency and achievement still form the major component of students' final course grade, but metalinguistic knowledge, engagement, agency, intent, and effort also play a part in assessment via the portfolio task. (p. 156)

One thread to correlate this study to the important consideration of validity in assessment is that this approach (e.g., portfolio assessment) encourages students to actively speak the target language in real-life situations, leading to a positive washback effect and a high level of authenticity. Also, the activity promotes the students' active engagement in real-life interactions as valuable sources for learning, emphasizing their concentration on particular learning instances within such interactions.

The seventh chapter, *The Role of Inscribed Object in a German Classroom-Based Paired Speaking Assessment: Does the Topic Card Help Elicit the Targeted Speaking and Interactional Competence?* by Katharina Kley, investigates the influence of a topic card on the interactions between paired students during a speaking assessment conducted in a classroom setting. The author discovered that most students were focused on the topic card, especially before transitioning to different topics. The participants would observe, indicate, or grasp the card before commencing a new subject, often by questioning with a query to obtain information. Students primarily utilized pre-shift tokens such as acknowledgments or exams to conclude the current topic before transitioning. The author proposes modifying the test methods by clearly defining the seating arrangements and limiting physical contact

with the card in order to provide a standardized approach to its usage. The grading rubric might be enhanced by establishing separate criteria for topic initiation and question-asking. In summary, this study emphasizes the need to carefully consider how the inclusion of inscribed objects in assessment tasks might impact test-takers' behavior and, as a result, the accuracy of the conclusions made from the assessment.

The eighth chapter, *First Language –Second Language Speaker Interaction: Affordances for Assessing Repair Practice*, by Katharina Kley, Silvia Kunitz, and Meng Yeh, specifically examines the methods employed for repair practices in two speaking tests conducted in a classroom setting. The purpose of these assessments is to evaluate the spoken language skills of college students studying Chinese as a foreign language. This study investigates the disparities in the utilization of other-initiated repair and other-directed word searches during interactions with a second-language speaker (L2S) and a native speaker (L1S) of Chinese. The researchers discovered nine specific actions individuals take to repair communication breakdowns begun by others, as well as five distinct ways in which individuals actively seek clarification or information from others. The most common practices observed in both test scenarios were requesting explanations with repeated reference to the problem's source, repeating the problem's source, explicitly asking for translations, and producing the next task to be done in English (Ahmadjavaheri & Zeraatpishe, 2020). The students demonstrated a higher level of involvement in repair activities, particularly when engaging in other-directed word searches while interacting with the L1S as opposed to the L2S. This indicates that the contact between the first language (L1) and the second language (L2) offered greater opportunities for commencing repair. The authors explain this phenomenon by referring to the linguistic epistemic imbalance between the learner and the L1S, where the learner perceives the L1S as having greater knowledge.

The ninth chapter, *Yardsticks for the Future of Language Assessment: Disclosing the Meaning of Measurement*, by Albert Weideman, scrutinizes the professional experiences concerning developing a theoretical basis for designing language assessments. The chapters' unorthodox stance toward the already existing principles can be traced through the statement that the effectiveness of a theory "may be due to the fact that the theory, when extended beyond its starting point, was turned into a rigid ideology, (p. 233)"; therefore, obtaining immunity to being contestable. As such, the argument's premise of the chapter is that applied linguistic designs, including language assessment designs, are constantly characterized by a technical approach. Within these contexts, the designer of the language intervention (whether it is a test, a language policy, or a language course) exercises control over the technical creativity, while the theoretical justification offers the analytical foundation for it. The latter facilitates the design process without imposing a scientific approach. The chapter has critically dealt with a lingual yardstick to address the meaningfulness of the design of the test and conceptualized it as "how informativity and transparency can be built into the design and administration of a test by making enough information available to test takers beforehand, especially a description of the components of the ability being tested, as well as a sample test" (p. 223). The chapter then examines the technological, social, economic, and ethical aspects that must be considered while creating responsible language tests. Albert Weideman asserts that progress has been achieved in several domains, including the significance of test design, social suitability and accessibility, technological usefulness and efficiency, alignment between concepts and tests, and accountability. Nevertheless, he recognizes that there are still obstacles to overcome, especially in guaranteeing integrity, benevolence, trustworthiness, and certainty in language test design.

The edited volume under review has contributed several innovative elements to language testing, assessment, and validity in theory and practice. The book has elucidated modifications in how we delineate language, both for assessment and personal development. It is necessary to carefully reconsider the framework we use to evaluate language proficiency and create educational programs to promote its development. Upon reevaluating the definition of language ability, another rational thought can be identified in the ongoing debate within validity theory. This debate encompasses the previously undisclosed notion of validity and its subsequent clarification through the concepts of construct, appropriateness, and social impact. This debate also leads to a broader understanding of validity and validation. In addition, the book has brought about a deeper awareness of ethical concerns, particularly with reference to the institutional systems that assess one's proficiency in an additional language. Owing to the fact that the institutional settings where tests are employed are

typically multilingual, these situations raise additional inquiries that compel researchers to develop assessments in a responsible manner. In this regard, language testing has faced difficulties in broadening the concepts of justice and fairness beyond their original interpretations and application in the field of language testing. Furthermore, the book has comprehensively addressed the subject matter commonly referred to as validity and validation critically. The writers have stated that their desire to develop assessments responsibly extends beyond the scope of what is presented in the book. The authors recognize that the advancements they witnessed and described in the book reveal an unfolding and disclosure of the meaning of the design, offering benchmarks for the future.

In summary, the book puts forward a concise and pragmatic approach to prevalent validity issues in practice and theory for researchers and scholars in the field of language testing and assessment. To this light, the raised topics have been thoroughly considered, not just as abstract notions but also with an effort to connect theory to real-world application whenever feasible. Furthermore, the book has critically unraveled a transformation in the definition of language ability, which necessitates a thorough reassessment of the concept itself, as well as the appropriateness and societal consequences of the assessments used to measure language ability or the design of courses aimed at enhancing it. The book has also landed a heightened awareness of ethical issues, particularly concerning the institutional frameworks where language ability is assessed.

In spite of the merits boasted by this edited volume, there seems to be some room left for improvements in some chapters of the book. Some chapters (e.g., chapters four, eight, and nine) can enjoy a more scrutinized proofreading to eliminate spelling and grammatical mistakes. In some cases, the authors can simplify the language to launch a more readable text, ensuring that complicated concepts are communicated clearly and effectively. This approach not only improves knowledge for a wider range of individuals, encompassing those who may not be familiar with the subject at hand, but also promotes active participation by minimizing obstacles to comprehension. Furthermore, in chapter three, the author's conclusion emphasizes the need for assessments that are both democratic and unbiased. However, the author does not provide concrete and practical measures that educators and politicians may use to achieve this goal. Offering specific instances of effective assessment methods would be advantageous. Chapter five focuses on student beliefs and perceptions of portfolio assessment, although it does not offer a comprehensive understanding of the many experiences and viewpoints within this subject area. The chapter briefly mentions the negative experiences of certain learners but does not delve into the relevant causes underlying these emotions. Last but not least, in order to improve chapter eight of the edited volume, the writers may consider offering more specific recommendations on how to refine the operationalization of repair in the current speaking test in the spirit of the findings. In addition, the researchers in this chapter could more deeply investigate the correlation between the repair scores of the students and their actual repair practices, as they have proposed for a prospective future study.

References

- Ahmadjavaheri, Z., & Zeraatpishe, M. (2020). The impact of construct-irrelevant factors on the validity of reading comprehension tests. *International Journal of Language Testing*, 10(1), 1-10.
- Davies, A., & Elder, C. (2005). Validity and validation in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 795–813). Lawrence Erlbaum Associates. <https://doi.org/10.4324/9781410612700>
- Eskildsen, S. W., PekarekDoehler, S., Piirainen-Marsh, A., & Hellermann, J. (2019). Introduction: On the complex ecology of language learning 'in the Wild'. In J. Hellermann, S. Eskildsen, S. PekarekDoehler & A. Piirainen-Marsh (Eds.), *Conversation analytic research on learning-in-action* (Vol. 38, pp. 1–21). Educational Linguistics, Springer. https://doi.org/10.1007/978-3-030-22165-2_1
- Kane, M. T., Kane, J., & Clouser, B. E. (2017). A validation framework for credentialing tests. In C. W. Buckendahl & S. Davis-Becker (Eds.), *Testing in the professions: Credentialing polices and practice* (pp. 20–41). Routledge.
- Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher*, 10(9), 9–20. <https://doi.org/10.3102/0013189X010009009>

- Schildt, L., Deygers, B. & Weideman, A. (2023). Language testers and their place in the policy web. *Language Testing*, 41(2), 338-356. <https://doi.org/10.1177/02655322231191133>
- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Longman.