International Journal of L2CT
LINGUISTICS, LITERATURE, CULTURE, AND TEACHING
Volume 1 Issue 3, 2026
DOI: https://doi.org/10.34050/l2ct.v1i3.45180

Indonesian B Teachers' Views on the Quality and Pedagogical Impact of IB Diploma Programme Paper 2 Assessment (2023–2025)

Adi Abdurahman
Universitas Nasional Jakarta, Indonesia
E-mail: adiabdurahman.yt@gmail.com
Tadjuddin Nur
Universitas Nasional Jakarta, Indonesia
E -mail: nurtadjuddin@gmail.com
Siti Tuti Alawiyah
Universitas Nasional Jakarta, Indonesia
E -mail: siti.tuti.alawiyah@civitas.unas.ac.id

ABSTRACT

This study investigates Indonesian B teachers' views on the quality and pedagogical impact of Paper 2 Reading and Listening assessments within the International Baccalaureate (IB) Diploma Programme during the 2023–2025 period. It aims to examine content alignment with the Language B Guide, evaluate assessment design quality, and explore the pedagogical implications of Paper 2 in classroom contexts. A qualitative approach was employed using semi-structured interviews with five experienced Indonesian B teachers, including two IB examiners. Data were analyzed using reflexive thematic analysis and subsequently organized within a SWOT framework to structure interpretive findings.
The results indicate that Paper 2 demonstrates strong alignment with the Language B Guide and supports higher-order thinking and intercultural reflection. However, teachers identified challenges related to linguistic complexity, accent variation in Listening, and fairness across diverse learner groups. The findings also reveal that the pedagogical impact of Paper 2 is mediated by teachers' assessment literacy and instructional practices. These findings contribute to the understanding of consequential validity and washback in international language assessment, highlighting the role of teachers' views as critical evidence in evaluating assessment quality.

Keywords: *Consequential validity, Indonesian B Paper 2, International Baccalaureate, teachers' views, washback.*

ARTICLE HISTORY
Published
April 21th 2026
Check for updates

1. Introduction

The International Baccalaureate (IB) Diploma Programme (DP) positions language learning as a vehicle for developing intercultural understanding, critical thinking, and global awareness (International Baccalaureate Organization, 2018). Within this framework, Language B courses are designed for second language learners to develop communicative competence and engage with diverse cultural perspectives (McKinley & Rose, 2022; Taguchi, 2026; Prihandoko et al., 2021; Youngsun et al., 2024).

Paper 2, the externally assessed component measuring receptive skills (Reading and Listening), contributes directly to students' final Diploma scores. As a high-stakes assessment, its quality and alignment with curricular objectives have substantial implications for teaching practice, learning outcomes, and interpretive fairness (Green, 2020; Kane, 2016). The Language B Guide specifies five prescribed themes (Identities, Experiences, Human Ingenuity, Social Organization, and Sharing the Planet), eight Language Acquisition Aims, and five Assessment Objectives (AO1–AO5) that collectively define the domain of competence to be assessed (International Baccalaureate Organization, 2018).

While quantitative analyses can establish structural alignment between test items and curricular specifications, they cannot fully capture the experiential and pedagogical dimensions of assessment quality. Teachers, as curriculum implementers and assessment users, occupy a unique position in evaluating how assessments function in classroom contexts (Cheng & Fox, 2017; Rahman et al., 2023; Karubaba et al., 2025; Adinda et al., 2025). Their professional judgments provide crucial evidence for consequential validity—the extent to which assessment use supports intended educational outcomes and avoids unintended negative consequences (Messick, 1995).

Despite the growing body of research on IB assessment quality (Badham et al., 2025; Badham & Furlong, 2022), empirical studies examining teachers' views on Indonesian B Paper 2 remain limited, particularly in relation to teacher-based evidence of consequential validity in Indonesian B assessment contexts. This gap is especially significant given the sociolinguistic diversity of Indonesian B learners, who range from native speakers to heritage learners and non-native learners, each bringing distinct linguistic and cultural resources to the assessment (Baker, 2022; Anggawirya et al., 2021; Said et al., 2021).

This study addresses this gap by investigating the following research question: How do Indonesian B teachers perceive the quality, alignment, and pedagogical impact of Paper 2 Reading and Listening assessments during the 2023–2025 period? Specifically, the study examines teachers' views on content alignment with the Language B Guide, assessment design quality, and difficulty calibration. It also explores the potential of Paper 2 to foster critical thinking and intercultural awareness, as well as

issues of fairness across diverse learner populations and the risk of exam-oriented instruction.

This study contributes to the literature by providing empirically grounded teachers' views as evidence of consequential validity and by offering insights into how assessment design interacts with pedagogical practice in IB Indonesian B contexts.

2. Literature Review

2.1 Consequential Validity and Teacher Views

Contemporary validity theory conceptualizes validity not as a fixed property of an instrument but as an evaluative judgment concerning the adequacy and appropriateness of inferences drawn from test scores (Kane, 2016; Messick, 1995). Within this framework, consequential validity addresses the social and pedagogical implications of assessment use, including its impact on teaching practices and learning outcomes.

More recent work has further refined this perspective. Aryadoust (2023) problematized the traditional conflation of scientific reasoning and ethical considerations within construct validity, arguing for a clearer distinction between empirical evidence of score meaning and the consequential basis of assessment use. This distinction is particularly relevant for high-stakes international assessments such as IB Paper 2. Chapelle (2021) extended this argument by emphasizing that validity claims must be supported through multiple sources of evidence, including stakeholder perspectives. Similarly, Ockey (2024) highlighted the importance of contextual factors—such as learner diversity and task characteristics—in shaping validity in second language assessment.

Taken together, these perspectives suggest that validity cannot be fully established through structural or statistical analysis alone but must incorporate experiential evidence from key stakeholders. In this context, teachers' views provide critical insight into how assessment design translates into classroom practice and learner experience. This study builds on these perspectives by examining teachers' views as evidence of consequential validity in IB Indonesian B assessment contexts.

2.2 Washback in Language Assessment

The concept of washback, first systematically examined by Alderson and Wall (1993), refers to the influence of testing on teaching and learning. Washback may be positive when assessments promote meaningful learning practices, or negative when they narrow instruction to test-oriented strategies (Hughes & Hughes, 2021).

In high-stakes contexts such as the IB Diploma Programme, washback effects tend to intensify due to the strong link between assessment outcomes and certification, university admission, and institutional reputation (Green, 2020). Recent research suggests that washback is mediated by multiple interacting factors, including teacher beliefs, institutional context, and assessment literacy (Rahman et al., 2023). Nguyen and

Nguyen (2026) further extended this concept through “washforward,” proposing that assessment design may shape long-term pedagogical orientations beyond immediate test preparation.

These perspectives indicate that washback is not a uniform effect but a dynamic process influenced by how teachers interpret and implement assessment demands. Therefore, examining teachers’ views provides valuable insight into how Paper 2 assessment may generate both intended and unintended pedagogical consequences.

2.3 Intercultural Communicative Competence and Assessment

Byram (2021) conceptualized intercultural communicative competence (ICC) as the ability to interpret, relate, and negotiate meaning across cultural contexts through critical reflection and empathy. In the IB Language B framework, assessment texts are expected to engage students with diverse cultural perspectives, thereby promoting international-mindedness as a central educational goal.

From a validity perspective, Kunnan (2017) argued that culturally valid assessments must ensure equal interpretive opportunities for all test-takers, regardless of their social or cultural backgrounds. This principle extends beyond statistical bias to include representational equity and accessibility of cultural content within assessment materials (Randall, 2021).

These perspectives suggest that language assessment should not only measure linguistic competence but also support intercultural understanding. Consequently, evaluating Paper 2 requires examining how assessment content and tasks facilitate intercultural engagement across diverse learner populations.

2.4 Fairness and Linguistic Diversity

In multilingual assessment contexts, fairness becomes a complex and multidimensional construct. Badham and Furlong (2022) demonstrated that even when constructs are theoretically comparable across languages, response patterns and academic approaches may vary across learner groups. Salaberry et al. (2023) further argued that contextualized validity must incorporate social, cultural, and ethical dimensions of test-takers’ experiences.

For Indonesian B, learner diversity—including native speakers, heritage learners, and non-native learners—introduces additional challenges for ensuring equitable assessment outcomes. These differences may influence not only linguistic performance but also access to cultural knowledge embedded in assessment materials.

This study addresses these issues by examining teachers’ views on fairness and accessibility in Paper 2, thereby contributing to a more nuanced understanding of validity and equity in multilingual assessment contexts.

3. Method

3.1 Research Design

This study adopts a qualitative research design using semi-structured interviews to explore Indonesian B teachers' views on Paper 2 assessments from 2023 to 2025. A qualitative approach was selected to capture the depth, nuance, and contextual complexity of professional judgments that cannot be fully addressed through quantitative methods (Braun & Clarke, 2022).

3.2 Participants

Five Indonesian B SL teachers from IB Diploma Programme schools in Jakarta, Bogor, and Denpasar participated in the study. Participants were selected through purposive sampling based on three criteria: (1) a minimum of five years of teaching experience, (2) direct involvement in Paper 2 instruction and preparation, and (3) familiarity with IB assessment standards. Two participants also served as IB examiners with experience in external assessment marking.

Participant profiles are presented in Table 1.

Table 1. Participant Profiles

Code	Gender	School Location	Teaching Experience	Educational Background
I1	Male	North Jakarta	>10 years	Bachelor's degree in English Education; Bachelor's degree in Indonesian Education; Master's degree in Cultural Studies; Master's degree in Indonesian Language Education
I2	Male	North Jakarta	>15 years	Bachelor's degree in English Education
I3	Male	South Jakarta	>10 years	Bachelor's degree in Indonesian Language and Literature; Master's degree in Linguistics
I4	Female	Bogor	>5 years	Bachelor's degree in Indonesian Language and Literature
I5	Female	Denpasar	>15 years	Bachelor's degree in Indonesian Literature; Master's degree in Language Teaching

The sample size was determined based on the principle of information power and data saturation (Hennink & Kaiser, 2022). Saturation was indicated during the fourth interview, with the fifth interview conducted to confirm the consistency and stability of emerging themes.

3.3 Data Collection

Data were collected through semi-structured interviews guided by a protocol structured around a SWOT framework (Strengths, Weaknesses, Opportunities, and Threats) to systematically explore evaluative dimensions of assessment quality and impact. The interview guide consisted of thirteen questions addressing content alignment, assessment design, difficulty calibration, pedagogical potential, cultural representation, fairness, and the risk of exam-oriented instruction.

All interviews were conducted in Indonesian, audio-recorded with informed consent, and transcribed verbatim. Participant identities were anonymized through coded identifiers to ensure confidentiality.

3.4 Data Analysis

Data were analyzed using reflexive thematic analysis (Braun & Clarke, 2022) following a systematic process of familiarization, open coding, category development, and theme construction. The analysis focused on identifying patterns related to content validity, cultural relevance, linguistic complexity, and pedagogical implications.

The resulting themes were subsequently organized within a SWOT framework to provide a structured interpretation of strengths, weaknesses, opportunities, and threats identified in teachers' views.

Trustworthiness was ensured through data triangulation, member checking, and audit trail documentation, following the criteria of credibility, dependability, and confirmability (Lincoln & Guba, 1985).

4. Results and Discussion

Thematic analysis of the interview data yielded seven interconnected themes that collectively characterize teachers' views on Paper 2 quality and impact. The findings are presented and discussed in relation to relevant theoretical frameworks.

Table 2. Summary of Themes, Key Findings, and Theoretical Links

Theme	Key Finding	Theoretical Link
Content Alignment	Broadly aligned with *Language B Guide*; minor thematic imbalance	Content validity (Green, 2020)

Assessment Design	Cognitive progression present; tension between complexity levels	Construct validity (Bachman & Palmer, 2010)
Difficulty Consistency	Generally stable; Listening 2024–2025 more demanding	Reliability (International Baccalaureate Organization, 2023, 2024, 2025)
Critical Thinking & ICC	Strong potential for higher-order thinking and cultural reflection	Washback (Alderson & Wall, 1993); ICC (Byram, 2021)
IB Learner Profile	Supports thinker, communicator, open-minded attributes	International-mindedness framework
Fairness	Unequal access across native/heritage/non-native populations	Fairness (Kunnan, 2017; Salaberry et al., 2023)
Exam-Oriented Risk	Dependent on teacher approach and assessment literacy	Washback and washforward (Nguyen & Nguyen, 2026)

Table 2 synthesizes the seven emergent themes and demonstrates that teachers' evaluations of Paper 2 extend beyond a purely technical appraisal of test design toward a more complex and layered understanding of assessment quality. Rather than depicting alignment, reliability, and fairness as discrete dimensions, the findings suggest that these constructs are perceived as interdependent and, at times, in tension with one another. For instance, while strong alignment with the Language B Guide supports claims of content validity, teachers simultaneously identify imbalances in thematic representation and variations in linguistic complexity that may introduce construct-irrelevant variance and affect equitable access to test performance.

This interplay highlights a critical insight: assessment quality in this context is not treated as a fixed or intrinsic property of the test but as an emergent construct shaped by the interaction between assessment design, learner diversity, and pedagogical mediation. In particular, the coexistence of positive affordances (e.g., promotion of higher-order thinking and intercultural reflection) and potential risks (e.g., exam-oriented instruction and differential accessibility) underscores the dual role of Paper 2 as both a measurement instrument and a pedagogical driver.

Moreover, the alignment between teachers' views and established theoretical frameworks—ranging from content and construct validity to fairness and washback—provides convergent evidence that strengthens the interpretive argument for consequential validity within an argument-based validation framework. At the same time, the presence of tensions across themes suggests that validity should be understood not as a binary condition but as a negotiated and context-sensitive process. Building on this integrative perspective, the following sections unpack each theme in detail to illustrate how these dimensions are enacted and interpreted within classroom practice.

4.1 Content Alignment with the *Language B Guide*

All five participants affirmed that Paper 2 generally aligns with the Language Acquisition Aims and Assessment Objectives specified in the *Language B Guide*. I1 noted that the variety of question types—including multiple choice, short answer, heading matching, and text-based justification—provides opportunities for students to demonstrate both explicit comprehension and implicit meaning interpretation. As I1 stated, the Paper 2 structure adequately reflects the AO framework outlined in the *Language B Guide*.

I2 observed that alignment is particularly evident in the analytical demands of certain items, noting that texts and questions encourage students to go beyond surface reading toward broader contextual analysis. I3 provided specific examples, citing texts about the Tanoker community (2023) and the story of a pedicab driver's child pursuing doctoral education (2024) as representations of how local contexts connect to global issues such as education and social mobility. I4 and I5 emphasized the interpretive dimension of assessment design, with I4 observing that students engage with meaning beyond retrieval and I5 noting that assessed skills feel authentic because texts represent Indonesian cultural realities.

However, participants expressed nuanced views regarding the proportionality of prescribed themes. While all five confirmed that the five major themes appeared across the 2023–2025 period, I3 and I4 noted that Social Organization and Sharing the Planet tended to dominate over Identity in the context of personal reflection. I2 did not view this imbalance as substantially problematic but suggested that thematic balance could be strengthened in future examination sessions.

These findings align with the concept of content validity (Green, 2020) but also extend it by demonstrating that alignment operates as a dynamic continuum shaped by thematic representation and contextual interpretation rather than a fixed binary condition. The teachers' recognition of both alignment and imbalance indicates that validity is not merely a structural property of test design but an interpretive construct mediated by classroom experience. This suggests that evaluating assessment quality requires integrating both content mapping and stakeholder-based evidence.

4.2 Assessment Design Quality and Critical Literacy

Beyond content alignment, teachers' views also reveal issues related to assessment design quality and cognitive demand. Participants highlighted the quality of item design and its relationship to critical literacy development. I2 and I3 identified a clear cognitive progression within Paper 2, moving from retrieval-level information identification to meaning interpretation and evaluative analysis. I4 emphasized that item types requiring text-based evidence, such as True/False with justification, strengthen critical literacy because students must demonstrate explicit textual support rather than rely on intuition. I5 observed that despite being comprehension-based, Paper 2 demands analytical and reflective processing, requiring students to assess relevance and meaning relationships within broader contexts. This graduated structure resonates with Hailikari et al. (2022) findings that constructive alignment between learning objectives, instructional activities, and assessment demands directly influences the quality of students' learning approaches, with misalignment promoting surface-level rather than deep engagement.

However, several participants raised critical observations about item complexity. I3 identified what he described as “trick questions” in synonym items that test not only language ability but also more complex reasoning. I1 and I4 noted the use of low-frequency vocabulary and relatively high register in some Reading texts, which they perceived as approaching Higher Level characteristics. I2, however, viewed variation in item difficulty as part of testing reading precision rather than a substantive problem.

Regarding the Listening component, I4 identified audio recordings with strong regional accents that posed difficulties for non-native students, while I5 emphasized that consistency in intonation and articulation is essential for maintaining assessment accessibility.

These divergent perspectives highlight a critical tension between cognitive and linguistic complexity, suggesting that construct validity in second language assessment is not only a matter of task design but also of how linguistic demands are calibrated relative to the intended proficiency level. This finding extends Bachman and Palmer’s (2010) framework by illustrating how construct-irrelevant variance may emerge when linguistic difficulty exceeds the target construct. It also underscores the need for ongoing calibration to balance depth of interpretation with accessibility.

4.3 Difficulty Consistency Across Years

In addition to design considerations, participants reflected on the consistency of difficulty across examination sessions. Participants demonstrated relatively balanced views regarding difficulty calibration across the 2023–2025 period. I2 described a graduated structure within each examination session, where students first encounter relatively straightforward informational texts before progressing to more explanatory and then interpretive passages. I3 observed general consistency across years but noted that Reading 2023 felt more challenging due to cultural topics (traditional arts, regional

songs) that may be unfamiliar to some Indonesian B students. I4 identified inter-component variation, noting that in some sessions Listening was more demanding than Reading due to the prevalence of implicit meaning requiring interpretation.

I1 specifically highlighted Listening concerns, observing that narration speed and articulation in some formal monologue segments increased noticeably compared to previous years, potentially affecting SL student comprehension. I5 noted the use of relatively high register and idiomatic vocabulary in some Reading texts, which she considered acceptable but requiring careful proportionality for students with diverse language backgrounds.

These observations corroborate findings in the Indonesian B Subject Reports (International Baccalaureate Organization, 2023, 2024, 2025), which noted that items with dual responses and complex clause structures represented the most challenging components for candidates. The convergence between teacher observations and official examiner reports strengthens the credibility of these findings through source triangulation.

Taken together, these findings suggest that difficulty consistency is not a fixed construct but a relative and context-dependent perception shaped by text familiarity, linguistic register, and modality differences. This triangulated evidence provides a robust basis for evaluating assessment reliability across examination sessions.

4.4 Critical Thinking, Inquiry, and Intercultural Awareness

Beyond technical aspects of assessment, teachers emphasized its pedagogical impact on critical thinking and intercultural awareness. All five participants affirmed that Paper 2 has strong potential for promoting critical thinking, intercultural awareness, and inquiry. I2 observed that the question structure inherently requires students to connect information, compare ideas, and understand issues within global contexts, directly training higher-order thinking skills. I3 explained that varied question wording and synonym usage encourage students to examine texts more carefully, building habits of reflective rather than hasty responding.

I4 stated that themes related to social and cultural issues open space for cross-cultural reflection, as students reading about specific social phenomena begin comparing these with their own life contexts. I5 added that although Paper 2 is receptive in nature, reading and listening processes involve evaluating information relevance, with students frequently motivated to question cause-and-effect relationships discussed in texts. I1 reported that in teaching practice, Paper 2 texts often serve as triggers for class discussions, with students raising follow-up questions that indicate emerging inquiry processes.

These findings have significant implications for understanding positive washback. In the framework of washback theory (Alderson & Wall, 1993; Cheng & Fox, 2017), the teachers' descriptions suggest that Paper 2 functions not merely as an evaluative

instrument but also as a catalyst for reflective thinking, critical analysis, and intercultural awareness development. This represents a form of positive washback where assessment design promotes meaningful learning practices beyond test preparation.

Furthermore, these observations connect to Byram's (2021) framework of ICC. When students engage with texts about local communities, social mobility, technological innovation, and environmental issues, and subsequently compare these contexts with their own experiences, they are exercising the interpretive and relational skills that constitute intercultural competence. The assessment thus serves as a pedagogical bridge between language comprehension and cultural understanding.

4.5 IB Learner Profile and International-Mindedness

These pedagogical effects further connect to broader IB educational goals reflected in the Learner Profile. Participants consistently associated Paper 2 with the development of IB Learner Profile attributes and international-mindedness. I1 observed that texts and audio materials addressing Indonesian cultural issues within global contexts enable students to view Indonesia from an international perspective. I2 identified alignment with Learner Profile attributes such as "thinker" and "communicator," noting that some items accommodate multiple interpretation possibilities, requiring students to communicate their reasoning.

I4 emphasized that IB-selected texts are not merely informative but reflective, training students to be open-minded and consider different viewpoints. I5 noted that Paper 2 provides space for students to examine social and cultural issues from multiple perspectives, ultimately strengthening global awareness. I3 observed that topics such as technology in remote areas and education as social mobility encourage students to understand social realities more broadly.

These findings indicate that Paper 2 contributes not only to language assessment but also to the broader educational mission of the IB by fostering dispositions aligned with the Learner Profile. The assessment design reinforces institutional values when content selection and task demands are aligned with broader philosophical goals, thereby extending the role of assessment beyond measurement toward identity and perspective formation.

4.6 Fairness of Access Across Learner Populations

However, the effectiveness of assessment must also be considered in relation to fairness across diverse learner populations. Teacher views on fairness revealed significant variation. I3 stated unequivocally that opportunities are not fully equal, particularly when fluent native speakers take the course without clear regulatory guidance, potentially influencing future examination difficulty calibration. I2 observed that opportunities depend heavily on student backgrounds, noting that certain non-native students may actually have advantages in understanding formal texts, while heritage learners may struggle with formal vocabulary rarely used in everyday communication.

I4 argued that Paper 2 is theoretically designed to measure comprehension-based skills, allowing background differences to be minimized through appropriate teaching strategies, while acknowledging differences in cultural context readiness among students. I5 noted that IB system flexibility allows students to choose pathways matching their abilities but emphasized the need for differentiated instructional strategies.

I1 viewed Indonesian B as a relatively inclusive learning space for non-native speakers while recognizing the need for pedagogical adaptations to ensure all students can demonstrate their abilities fairly.

These findings resonate with Kunnan's (2017) framework of assessment fairness, which extends beyond technical equality to encompass equitable opportunities for demonstrating relevant competence. The diversity of perspectives suggests that fairness in Indonesian B assessment operates at the intersection of assessment design, linguistic background, and pedagogical mediation. When heritage learners possess strong cultural intuition but limited formal register knowledge, while non-native learners demonstrate stronger academic language skills, performance reflects not only receptive competence but also the distribution of linguistic experience—a form of construct-relevant variance that requires pedagogical differentiation rather than assessment redesign (Salaberry et al., 2023).

4.7 Risks of Exam-Oriented Instruction

Finally, participants reflected on the broader instructional consequences of Paper 2, particularly the risk of exam-oriented teaching. On this issue, participants expressed divergent views. I3 identified the potential for instruction to become exam-oriented when teachers overemphasize answer strategies over holistic language development. Conversely, I2 argued that Paper 2 promotes exploratory orientation because the texts used open space for discussion and broader understanding. I1 and I5 acknowledged both dimensions, recognizing that Paper 2's assessment weight can trigger exam focus while noting that this outcome depends heavily on teachers' instructional approaches.

Regarding external factors, I2 highlighted school support infrastructure and teacher training as critical factors in maintaining alignment between instruction and Paper 2 content. I4 added that the availability of authentic resources and sufficient time for receptive skill development also present challenges. I5 noted that format changes or IB curriculum developments require ongoing information updates to maintain alignment.

These findings illustrate the dynamic nature of washback as theorized by Nguyen and Nguyen (2026), who proposed that assessment design can shape long-term pedagogical orientations ("washforward"). The teachers' observations suggest that Paper 2's pedagogical potential is mediated by assessment literacy—teachers who understand the communicative and reflective intentions behind assessment design are more likely to generate positive washback, while those who focus primarily on scoring mechanics may inadvertently narrow instruction. This finding is consistent with Gokturk

Saglam and Tsagari (2022) who demonstrated that teacher and student perceptions of assessment consequences significantly shape the direction and intensity of washback, and that divergent stakeholder views can coexist regarding the same assessment instrument.

This finding has practical implications: ensuring positive washback from Paper 2 requires not only well-designed assessment items but also professional development that strengthens teachers' capacity to interpret assessment purposes and translate them into meaningful classroom practice (Rahman et al., 2023).

As Xu and Liu (2024) argued, teacher assessment literacy encompasses not only technical knowledge of testing but also the capacity to interpret assessment purposes and translate them into pedagogically sound classroom practices—a competency that directly mediates washback outcomes in high-stakes contexts.

5. Conclusion

This study has examined Indonesian B teachers' views on the quality and pedagogical impact of Paper 2 Reading and Listening assessments during the 2023–2025 period. The findings indicate that Paper 2 demonstrates strong alignment with the Language B Guide and effectively supports the development of receptive competence across retrieval, interpretation, and evaluation levels. In addition, the assessment shows considerable potential to promote critical thinking, intercultural awareness, and the development of IB Learner Profile attributes.

However, the study also identifies several critical tensions. First, the boundary between cognitive complexity and linguistic complexity represents a key challenge in maintaining construct validity, as excessive linguistic demands may introduce construct-irrelevant variance. Second, fairness across native, heritage, and non-native learner populations emerges as a context-dependent issue that requires pedagogical mediation rather than uniform assessment design. Third, the direction and intensity of washback are strongly influenced by teachers' assessment literacy and instructional practices.

The study contributes to the literature on consequential validity in international language assessment by demonstrating that assessment quality cannot be fully evaluated through structural alignment alone. Instead, teachers' views provide essential empirical evidence of how assessment design is interpreted, enacted, and experienced in classroom contexts.

Future research should extend this work by incorporating student views and examining classroom practices through observational and mixed-method designs to further strengthen the evidential basis for validity claims in IB language assessment. These findings have important implications for assessment design in multilingual contexts, particularly in ensuring equitable access and meaningful learning across diverse learner populations.

References

1. Adinda, R., Sosrohadi, S., Syafitri, B. A., & Andini, C. (2025). Cognitive And Cultural Barriers In Synonim Acquisition: A Psycolinguistic Study Of Indonesian Learners Of Korean. *TPM–Testing, Psychometrics, Methodology in Applied Psychology*, *32*(4), 881-888.
2. Anggawirya, A. M., Prihandoko, L. A., & Rahman, F. (2021, December). Teacher's role on teaching English during pandemic in a blended classroom. In *International Joined Conference on Social Science (ICSS 2021)* (pp. 458-463). Atlantis Press.
3. Alderson, J. C., & Wall, D. (1993). Does washback exist? Applied Linguistics, 14(2), 115–129. https://doi.org/10.1093/applin/14.2.115
4. Aryadoust, V. (2023). The vexing problem of validity and the future of second language assessment. Language Testing, 40(1), 8–14. https://doi.org/10.1177/02655322221125204
5. Bachman, L., & Palmer, A. (2010). Language assessment in practice. Oxford University Press.
6. Badham, L., & Furlong, A. (2022). Summative assessments in a multilingual context: What comparative judgment reveals about comparability across different languages in Literature. International Journal of Testing, 23(2), 111–134. https://doi.org/10.1080/15305058.2022.2149536
7. Badham, L., Meadows, M., & Baird, J. A. (2025). Construct comparability and the limits of post hoc modeling: insights from International Baccalaureate multi-language assessments. Frontiers in Education, 10. https://doi.org/10.3389/feduc.2025.1616879
8. Baker, W. (2022). Intercultural and transcultural awareness in language teaching. Cambridge University Press. https://doi.org/10.1017/9781108874120
9. Braun, V., & Clarke, V. (2022). Conceptual and design thinking for thematic analysis. Qualitative Psychology, 9(1), 3–26. https://doi.org/10.1037/qup0000196
10. Byram, M. (2021). Teaching and assessing intercultural communicative competence: Revisited (2nd ed.). Multilingual Matters.
11. Chapelle, C. A. (2021). Argument-based validation in testing and assessment. SAGE Publications. https://doi.org/10.4135/9781071878811
12. Cheng, L., & Fox, J. (2017). Assessment in the language classroom. Bloomsbury Publishing.
13. Gokturk Saglam, A. L., & Tsagari, D. (2022). Evaluating perceptions towards the consequential validity of integrated language proficiency assessment. Languages, 7(1), 65. https://doi.org/10.3390/languages7010065
14. Green, A. (2020). Exploring language assessment and testing: Language in action (2nd ed.). Routledge. https://doi.org/10.4324/9781003105794
15. Hailikari, T., Virtanen, V., Vesalainen, M., & Postareff, L. (2022). Student perspectives on how different elements of constructive alignment support active learning. Active Learning in Higher Education, 23(3), 217–231. https://doi.org/10.1177/1469787421989160

16. Hennink, M., & Kaiser, B. N. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests. Social Science and Medicine, 292. https://doi.org/10.1016/j.socscimed.2021.114523
17. Hughes, A., & Hughes, J. (2021). Testing for language teachers (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781009024723
18. International Baccalaureate Organization. (2018). Language B guide: First assessment 2020. www.ibo.org
19. International Baccalaureate Organization. (2023). May 2023 subject report Indonesian B.
20. International Baccalaureate Organization. (2024). May 2024 subject report Indonesian B.
21. International Baccalaureate Organization. (2025). May 2025 subject report Indonesian B.
22. Kane, M. T. (2016). Validity as the evaluation of the claims based on test scores. Assessment in Education: Principles, Policy & Practice, 23(2), 309–311. https://doi.org/10.1080/0969594X.2016.1156645
23. Karubaba, S., & Rahman, F. (2025). Code-Switching and Code-Mixing in Indonesian EFL Classrooms: Teacher-Student Interactions in North Biak. *Dialectica Online Publishing Journal*, *1*(1), 107-115.
24. Kunnan, A. J. (2017). Evaluating language assessments. Routledge.
25. Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. SAGE Publications.
26. McKinley, J., & Rose, H. (2022). English language teaching and English-medium instruction: Putting research into practice. Journal of English-Medium Instruction, 1(1), 85–104. https://doi.org/10.1075/jemi.21026.mck
27. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741
28. Nguyen, T. L., & Nguyen, T. N. Q. (2026). Washback, washforward, and the possibility of a co-constructive harmony of test use and test development. Language Testing in Asia, 16(1). https://doi.org/10.1186/s40468-025-00406-4
29. Ockey, G. J. (2024). Introducing Second Language Assessment. Cambridge University Press.
30. Prihandoko, L. A., Anggawirya, A. M., & Rahman, F. (2021, December). Students' perceptions towards autonomous learners concept in academic writing classes: Sequential mixed-method. In *International Joined Conference on Social Science (ICSS 2021)* (pp. 487-491). Atlantis Press.
31. Rahman, K. A., Rukanuddin, M., Chowdhury, M. S. Y., Ahmed, S., & Seraj, P. M. I. (2023). Recognizing stakeholders and factors mediating washback in language testing. Education Research International, 2023(2023), 5548723. https://doi.org/10.1155/2023/5548723
32. Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. Educational Measurement: Issues and Practice, 40(4), 82–90. https://doi.org/10.1111/emip.12429

33. Salaberry, M. R., Weideman, A., & Hsu, W.-L. (2023). Ethics and context in second language testing: Rethinking validity in theory and practice (M. R. Salaberry, A. Weideman, & W.-L. Hsu, Eds.; 1st ed.). Routledge. https://doi.org/10.4324/9781003384922
34. Said, M. M., Rita, F., Weda, S., & Rahman, F. (2021). English language performance development through extracurricular activities at Faculty of Teacher Training and Education Tadulako University Palu. *PalArch's Journal of Archaelogy of Egypt/Egyptology*.
35. Taguchi, N. (2026). Development of intercultural communicative competence in an English-medium university in Japan. TESOL Quarterly, 60, 321–348. https://doi.org/10.1002/tesq.70038
36. Xu, Y., & Liu, Y. (2024). Language assessment literacy for teachers: A systematic review. In Z. Tajeddin & T. S. Farrell (Eds.), Handbook of language teacher education. Springer. https://doi.org/10.1007/978-3-031-43208-8_16-1
37. Youngsun, K., Sosrohadi, S., Andini, C., Jung, S., Yookyung, K., & Jae, P. K. (2024). Cultivating Gratitude: Essential Korean Thankfulness Phrases for Indonesian Learners. *ELS Journal on Interdisciplinary Studies in Humanities*, *7*(2), 248-253.