

Bridging HOTS and character education in reading assessments: A study on validity, reliability, and Pancasila integration

Imelda Gozali¹, Faizatul Istiqomah^{2*}, Ly Gia Ngan Vo³

¹English Language Education Study Program, Teacher Education Faculty, Universitas Katolik Widya Mandala Surabaya, Indonesia, ²English Language Education Department, Faculty of Education and Teacher Training, Universitas Islam Negeri Sayyid Ali Rahmatullah Tulungagung, Indonesia, ³English Linguistics and Literature Department, University of Social Sciences and Humanities (USSH-VNUHCM) – HCMC National University, Vietnam

*) Corresponding Author, email: faizatulistiqomah79@gmail.com

ABSTRACT

Very little research on developing reading assessments has systematically integrated higher-order thinking skills (HOTS) and character building in an original, valid, and reliable reading comprehension test in English. This study addresses that gap by modeling the development of such an instrument through a research and development (R&D) design comprising four stages: preliminary study, development, evaluation-revision, and pre-implementation. Within these stages, a test blueprint based on reading sub-skills and HOTS indicators was created, and Pancasila student profile guided the selection of the four reading texts. The 40 multiple-choice questions were analyzed for readability, reliability, item difficulty, and item discrimination. Participants included 12 undergraduate students enrolled in an English reading course, with feedback from the course lecturer providing face validity through a reviewer checklist. The try-out results showed that the test passed the face validity and reliability check, and the majority of the questions were found to have moderate difficulty (48%) and satisfactorily discriminating (38%). To improve item quality, several questions were revised and their options refined. This study contributes a replicable model for teachers and teacher educators, demonstrating how classroom-based reading assessments can be designed to foster evaluative thinking and character formation. While limited by its small sample size and less stringent item revision criteria, the findings provide practical guidance for assessment design and inform teacher training and professional development programs. Future research could implement the revised test in larger-scale contexts, gather students' perspectives, and examine the relationship between HOTS levels and item difficulty to refine assessment practices further.

Keywords: character education; HOTS; Pancasila; reading comprehension; R&D method

First Received:	Revised:	Accepted:	Published:
December 12 th , 2025	March 20 th , 2026	March 21 st , 2026	March 28 th , 2026

How to cite (in APA style):

Gozali, I., Istiqomah, F., & Vo, L. G. N. (2026). Bridging HOTS and character education in reading assessments: A study on validity, reliability, and Pancasila integration. *Journal of Research on English and Language Learning (J-REaLL)*, 7(1), 247–270. <https://doi.org/10.33474/j-reall.v7i1.24707>

INTRODUCTION

The challenges teachers face when having to develop their own, original assessment, are all too well-known (Atjonen, 2014). Overworked teachers lack the time and resources to design assessments that both align with the curriculum and foster learning (DeLuca & Bellara, 2013). As a result, teachers at times resorted to browsing for available test materials on the internet or relying on generative Artificial Intelligence (Bower et al., 2024). This often results in problems of dishonesty when students are under-supervised during online examinations, as frequently happened during the COVID-19 pandemic (Newton & Essex, 2024). Hence, there is a dire need for teachers to be able to develop an original, custom-made, valid, and reliable test to be used as a classroom-based, formal assessment.

On the other hand, teacher training institutions may not have sufficiently provided the needed foundational knowledge in classroom assessment design. A recent study measuring preservice teacher assessment competency revealed non-optimal results in the practical assessment skill and technological integration in the assessment design (Hoang et al., 2025). In addition, although preservice teachers might have acquired sound conception and base knowledge of assessment, they still lack the skill to collaborate with others to create meaningful assessments (Atjonen et al., 2022). Taken together, the insufficiency in the preservice teacher training with regard to assessment design might lead to a persistent tension when the preservice teachers practice their profession, especially when they need to align both the micro context (students' characteristics) and the macro level (educational policy) (Pastore, 2023).

The above description holds true in the milieu of English as a foreign language (EFL) teaching and learning, particularly in Reading classes. In the EFL context, teachers frequently need to develop their own reading material to adapt the content to the local culture and thus facilitate text understanding (Oktarina et al., 2022). Consequently, teachers will need to develop their own reading comprehension test battery, which suits the text content and the students' characteristics (Arshad, 2020). However, due to the insufficient coverage of assessment design training in the preservice teacher education, as previously mentioned, in-service EFL teachers were frequently found unaware of the standards for developing classroom-based assessment (Hanafi, 2016). In addition, they may not be familiar with the need to address the various reading subskills (Brown & Lee, 2015) which constitute a good reading comprehension test.

In the context of Indonesia, where this research is conducted, the abovementioned problem is further compounded by the need to align the test content with the continuous curriculum reform and mandate (Faradella et al., 2024). The prevailing Kurikulum Merdeka (Emancipated Curriculum) emphasizes the need to foster crucial thinking skills, embodied by the higher-order thinking skill (HOTS) (Rahmiati & Emaliana, 2017). As such, Indonesian EFL teachers are also exhorted to incorporate HOTS in their reading and assessment materials. Despite the numerous studies investigating the incorporation of HOTS into reading assessments, teachers are still faced with formidable challenges, ranging from the lack of resources to students' diverse cognitive skills (Tyas et al., 2019). Hence, there is still a need for teachers to develop the skill of constructing their own reading tests that, at the same time, foster students' critical thinking skills.

At the same time, there is a growing concern about Indonesian students' increasingly negative behaviour, as indicated by rising cases of bullying, drug consumption, and violent acts (Oktriani et al., 2023; Zakso et al., 2022). Thus, Merdeka Curriculum also emphasizes character education, to be imparted either as a distinct subject or integrated in all courses. To this end, the Indonesian government has been advocating the integration of Profil Pelajar Pancasila, or the Pancasila student profile, into the school's curriculum. Pancasila is the state's ideological values comprising five principles, namely (1) belief in the one God, (2) a just and civilized humanity, (3) national unity, (4) democracy led by people's representatives, and (5) social justice. The Pancasila student profile itself consists of six components, namely piety, diversity, solidarity, autonomy, critical thinking, and creativity, together with the attendant key elements, as detailed on the ministry website (<https://cerdasberkarakter.kemdikbud.go.id/profil-pelajar-pancasila/>).

While this framework provides the backbone for values integration in theory, research indicates that its implementation significantly shapes teachers' assessment practices. For preservice teachers, the policy requires assessment literacy that embeds moral and civic values into test design, yet studies show that teacher education programs often provide insufficient training in this area as mentioned previously (Hoang et al., 2025). For in-service teachers, the policy adds complexity to classroom assessment, as they must balance HOTS-oriented items with value-based content, often without adequate resources or professional development support (Haniah et al., 2020). Thus, while the Pancasila Student Profile offers a normative vision of holistic education, its practical application in EFL reading assessment remains a challenge, underscoring the need for research-driven models that demonstrate how HOTS and character education can be simultaneously embedded in valid and reliable instruments.

In sum, EFL teachers face the dual challenge of incorporating character education, in the form of the Pancasila Student Profile, and HOTS into their teaching and assessment materials. However, few studies have developed classroom-based reading comprehension instruments that simultaneously integrate HOTS with Pancasila values in a valid and reliable way. This absence is critical, as teachers perceive the ability to design such assessments as essential for fostering evaluative thinking and character formation (Ratrisari et al., 2024). Addressing this gap, therefore, is necessary to align classroom practice with the goals of the Merdeka Curriculum and the Pancasila Student Profile. Thus, this study aims to address this challenge by providing a model for developing a valid and reliable original, custom-made reading comprehension assessment, as well as producing the final, ready-made instrument that covers reading sub-skills, HOTS, and the good values that strengthen students' character. Without such an effort, teachers may continue to depend on LOTS-heavy assessments that fail to cultivate evaluative and critical thinking skills, thereby weakening the intended outcomes of the Merdeka Curriculum and the Pancasila Student Profile.

Therefore, the research is guided by the following questions:

1. What are the stages involved in developing an original English reading comprehension assessment that incorporates Pancasila-based character-building values and HOTS?
2. What steps can be taken to ensure that the developed reading assessment is valid and reliable?

LITERATURE REVIEW

Principles and practice of developing a reading assessment

Designing a classroom language test begins with a clear purpose and measurable objectives derived from curriculum goals (Brown & Lee, 2015). Test specifications outline the skills to be assessed, item types, and tasks, including details such as item distribution and time allocation. For reading assessment, teachers must attend to multiple components to ensure alignment with classroom objectives. Brown and Lee (2015) identified micro-skills (e.g., word recognition, grammatical class recognition, cohesive device recognition) and macro-skills (e.g., inference, identifying main ideas, distinguishing literal and implied meaning). Grabe and Jiang (2013) further emphasized twelve evidence-based skills, including vocabulary, grammar, and main idea formulation, as predictors of successful reading comprehension. Among reading test formats, multiple-choice questions remain the most widely used due to ease of administration, objectivity in scoring, and practicality for large groups (Butler, 2018; Özdemir & Akyol, 2019). Brown and Lee (2015) stressed that items should reflect strategies such as scanning, skimming, inferencing, and discourse marker use. Magliano et al. (2007) proposed a product-and-process framework for multiple-choice items, distinguishing local textbase, global textbase, and inference questions, thereby enhancing the validity of comprehension assessment. In addition, validity is central to classroom-based assessment. Content validity ensures that test items measure the intended skills (Sackett, 2012), while criterion-related validity examines alignment with learning objectives (Brown & Lee, 2015). Face validity, reflecting stakeholder perceptions of test appropriateness, can be established through expert or student feedback (Sato & Ikeda, 2015). Together, these forms of validity safeguard the authenticity and relevance of teacher-designed assessments (Afflerbach, 2016).

Besides validity, reliability ensures consistency in test administration and scoring (Brown & Lee, 2015). For receptive skills like reading, internal reliability is crucial, often measured through item analysis and internal consistency (Weir, 2005). In addition, item difficulty (p) indicates the proportion of correct responses, ideally ranging between 0.2 and 0.8 (Khairani & Shamsuddin, 2016). Items that are too easy or too difficult reduce reliability (Bachman, 2010). Item discrimination (D) measures how well an item distinguishes higher-ability from lower-ability students, with values above 0.2 considered effective (Miller et al., 2009). Internal consistency, measured by Cronbach's alpha or KR-20, further confirms the dependability of test scores (Nitko & Brookhart, 2014). Several studies have developed and validated reading assessments using these principles. Seifert and Paleczek (2021) created a digital tool for Austrian students, analyzing item difficulty, discrimination, and reliability. Özdemir and Akyol (2019) designed a test for fourth graders, computing discrimination, difficulty, and KR-20 reliability. Hanafi (2016) developed a test for undergraduates, revising items based on validity and reliability analyses. Ningrum and Sudarwati (2022) constructed a test for a Critical Reading course, though low student scores suggested limitations in sample size. These studies demonstrate the importance of validity and reliability in reading assessment design. Building on this foundation, the present study develops an English reading comprehension test that integrates HOTS and Pancasila-based character values, addressing both methodological rigor and curricular demands.

HOTS-based assessment

In determining the cognitive level of tests, teachers are commonly guided by taxonomies of thinking processes, such as Bloom's. Coined by Benjamin Bloom in 1956, the taxonomy comprises six cognitive levels arranged as a pyramid: *knowledge, comprehension, application, analysis, synthesis, and evaluation* (Bloom et al., 1956 in Krathwohl, 2002). The first three levels are known as lower-order thinking skills (LOTS), while the top three represent higher-order thinking skills (HOTS). Anderson and Krathwohl later refined the taxonomy into *remember, understand, apply, analyze, evaluate, and create* (Krathwohl, 2002). In Indonesia, the Merdeka Curriculum mandates the incorporation of HOTS in both teaching and assessment (Faradella et al., 2024). For assessment, HOTS integration is typically achieved by employing operational verbs from the *analyze, evaluate, and create* levels in question items. However, teachers face significant challenges in designing assessments that integrate HOTS while meeting course objectives (Amali et al., 2022; Rachmawati & Purwati, 2021). These challenges stem from varied teacher perceptions of HOTS (Setiawan et al., 2019), limited time and resources, and differences in students' cognitive competence (Tyas et al., 2019). Several studies have examined the cognitive levels of questions in reading assessments. Khorsand (2009), for example, investigated assessments designed by Iranian EFL teachers and found that most questions (54.21%) were at the *knowledge* level, the lowest rung of Bloom's taxonomy. Similar studies in Indonesia revealed comparable outcomes, with LOTS questions dominating over HOTS (Ardiana & As Sabiq, 2020; Putra & Abdullah, 2019). These findings highlight the need for teachers to construct assessments at higher cognitive levels to foster critical thinking. To address these challenges, several studies have modelled the development of HOTS-based questions in assessment instruments. Rahmiati and Emaliana (2017) designed a reading comprehension test for ESP students, mapping each item onto Bloom's six cognitive domains to ensure balanced distribution of LOTS and HOTS. The draft test was subjected to expert judgment, readability checks, and item analysis, with revisions made to overly easy or difficult items. In addition, Setiawan et al. (2019) employed the ADDIE design to generate HOTS-based assessment instruments for fifth-grade students across various subjects. Cognitive levels were determined according to basic competencies, and blueprints were constructed accordingly. The resulting questions were validated by experts, tested for validity, and evaluated for practicality. Out of 25 drafted items, 17 were found valid, reliable, and practical. These studies demonstrate both the potential and the challenges of HOTS integration, underscoring the importance of developing models that guide teachers in constructing effective HOTS-based reading assessments.

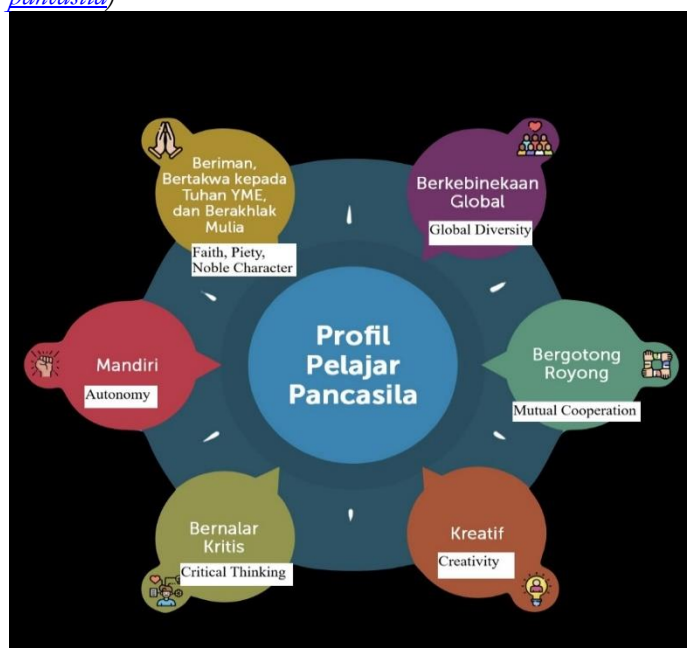
Character-based education: Infusing Pancasila characters into the curriculum

Besides developing students' cognitive and academic ability, the purpose of education is also instilling upright character in the youth, through the role models from the educators, promotion of intrinsic motivation, and direct teaching, among several others (Berkowitz, 2012). Thus, character education must be infused in all courses. Character education goes beyond moral lessons or simple teaching of what is right or wrong. Instead, it inculcates good habits in the pupils, in all the knowledge domains of cognitive (knowing the good things), affective (willingness to do the good things) and psychomotor (actually doing the good things) (Sudarwati et al., 2021), which ultimately brings about the phronesis or the students' knowledge of wisdom (Hidayat et al., 2021). Character education can also be integrated into language teaching, as shown by the study of Nenohai et al. (2021), who investigated the implementation of character education in the EFL classroom of a Junior high school in Indonesia, applying the 2013 curriculum. The research found that character education values were not internalized in a similar manner; rather, they were shaped by teachers' inclusive perspective. In addition, the application of character education was found to enhance students' critical thinking skills (Nenohai, 2021).

In its Strategic Plan for the years 2020-2024, the Indonesian Ministry of Education and Culture has identified six priority characters that will enable the young generation to succeed in the future work environment. The six characters are encapsulated under the title of Profil Pelajar Pancasila ("Profil") or the Pancasila Student Profile. It requires students to develop their characters, having global diversity, mutual cooperation, creativity, critical thinking, autonomy, and faith in the One God, together with nobility of character. A graphic representation of the Profile is given as Figure 1.

Figure 1

Graphic Representation of Pancasila Student Profile (from <https://cerdasberkarakter.kemdikbud.go.id/profil-pelajar-pancasila>)



In response to the call from the government to develop the Pancasila profile in students, several studies have shown the integration of the five ideological principles of Pancasila into primary to tertiary curriculum in Indonesia. Most of the integration was in the form of teaching material (Ahmadi et al., 2021), learning materials such as a learning book (Arifin et al., 2020; Sari & Ariswan, 2021), learning media (Wuryandani et al., 2020), and modules (Sari & Ariswan, 2021; Wijayanti et al., 2022). The integration was not only for Pancasila education (Abdulkarim et al., 2020), but also other non-civic education subjects. Ahmadi et al. (2021) integrated Pancasila values into their comics for primary

school students learning Mathematics, whereas [Arifin et al. \(2020\)](#)'s study found that incorporating Pancasila value to Quick on The Draw Model assisted by the physics learning book can improve students' critical thinking skills. In addition, [Sari et al. \(2020\)](#)'s survey showed that the integration of Pancasila into the optical devices learning book was practicable and appropriate for senior high school students.

A few studies were found to incorporate the Pancasila Student Profile into teaching and learning materials. [Brata et al. \(2022\)](#)'s study, for example, revealed a significant correlation between the Pancasila Student Profile integration into the lesson plan and students' attitude construction. Their study indicated that students were able to demonstrate positive attitudes as indicated by the six components of the Pancasila Student Profile. In addition, [Shofa \(2021\)](#) investigated the implementation of Pancasila values in history class through project-based learning by making a replica of stone age culture. The results showed that the learning approach was able to scaffold the dimension of Pancasila student profile.

From the aforementioned studies to the best of the researchers' knowledge, the integration of Pancasila Student Profile values into assessment has not yet been investigated, especially for reading comprehension in English lessons. A study closely related to ours is that of [Ratrisari et al. \(2024\)](#), who examined Indonesian Language teachers' perceptions of integrating HOTS and Pancasila values into a summative assessment guideline for grade VII students. Similar to our study, [Ratrisari et al. \(2024\)](#) emphasized the importance of embedding HOTS and Pancasila values into reading-viewing skills through descriptive and narrative texts. However, their work was limited to teacher perceptions and did not extend to the actual development or validation of test items. In contrast, the present study advances this line of inquiry by systematically designing, revising, and validating an English reading comprehension test that integrates HOTS with multiple Pancasila values. This relevance underscores the originality of our contribution, as it moves beyond conceptual recommendations to provide a practical, replicable model for classroom-based assessment aligned with the Merdeka Curriculum.

METHOD

In educational research, the choice of design must balance theoretical rigor with contextual relevance, ensuring that the study setting is adequately represented. Situated in an Indonesian EFL program, this study acknowledges limitations such as the relatively small sample size and the constraints of classroom-based implementation, which may affect generalizability. These limitations, however, also highlight directions for future research, including larger-scale trials and replication across institutions. The present design was selected to provide a systematic and transparent account of how a reading comprehension assessment can be developed and validated while integrating HOTS and Pancasila-based character values. Accordingly, a research and development (R&D) approach was employed, as it allows iterative refinement of test items and ensures both validity and reliability in the resulting instrument.

R&D design as posited by [Borg and Gall \(1983, in Gustiani, 2019\)](#) is deemed to be the most apt for the purpose of this study, as R&D aims to develop products or procedures which are evaluated and tested until the desired criteria are met. While the original Borg and Gall (1983) model consists of 10 stages, several subsequent works modified and simplified the steps in accordance with the needs and the context of individual research ([Gustiani, 2019](#)). Based on the review studies employing various adaptations of the [Borg and Gall \(1983\)](#) model, the work of [Rofii et al. \(2018\)](#) is seen to suit the concept of this present study. This simplified [Borg and Gall \(1983\)](#) then comprises four steps: preliminary study, development, evaluation-revision, and implementation.

In line with the first purpose of this study, which is to model the development of a valid and reliable reading assessment instrument, a suitable reading course is needed to serve as the context of this research. It was decided to develop the reading assessment in one of the ongoing courses in an English Language Education Study Program, Teacher Education Faculty, in a private university in Surabaya, Indonesia. After a discussion with the study program's management, the Reading Fiction course was selected. This is a compulsory subject for second-semester students, worth 2 credit hours.

Ethical clearance for this study was obtained from the Research and Community Service Institute of the first authors' university, ensuring that all procedures complied with institutional guidelines for research involving human participants. This study involved twelve undergraduate students (5 males, 7 females) who were taking the Reading Fiction course. We also sought the cooperation of the Reading Fiction instructor, who is a faculty member of the same university. The lecturer assisted us in providing the course syllabus and materials, as well as to concretize the date for the assessment try-out with the students. Familiarising ourselves with the course syllabus is essential in ensuring the content validity of classroom-based assessment in general, and the assessment that we are developing in particular.

FINDINGS

Data collection – the preliminary study and development stages

The data collection process follows the R&D design as previously described. In the preliminary study stage, an analysis is conducted to reveal the practices, norms, and behaviours of the studied settings. In our study, it includes examining the prevailing syllabus, course and evaluation materials, as well as the assessment practices in the Reading Fiction course of the English Language Education Study Program. The syllabus was specifically examined in the course description and the expected learning outcomes section, which serve as input for the test specification. The analysis of the Reading Fiction syllabus revealed that the course description and objective explicitly mention the reading subskills and HOTS components. Besides, the topic of the texts for this course also contains character-building elements, such as “love”, “freedom”, and “making a difference.” This finding contributes to the content validity of the assessment that is being developed.

After getting familiarized with the Reading Fiction course, its characteristics and needs, the next step is drafting the guidelines for the reading text selection within the Development phase. As the *Profil Pelajar Pancasila* (“Profil”) was chosen as the model of good character, the criteria for finding the reading materials follow the components of the *Profil*. Table 1 shows the *Profil*'s components and key elements as translated from the Ministry of Education website.

Table 1

The Profile of Pancasila Students and its Key Elements

No	Profile Components	Key Elements
1	Faith and Piety in the One God and Nobility of Character (<i>beriman, bertakwa, berakhlak mulia</i>)	Religious morality, personal morality, inter-personal morality, morality to nature, and morality to the state
2	Global Diversity (<i>berkebinekaan global</i>)	Knowledge and appreciation of culture, intercultural communication, reflection, and responsibility
3	Mutual cooperation (<i>gotong royong</i>)	Collaboration, care, sharing
4	Autonomy (<i>mandiri</i>)	Self-awareness and self-regulation
5	Critical Thinking (<i>bernalar kritis</i>)	Getting and processing information, analyzing and evaluating thinking, reflection on thoughts and thought processes, decision making.
6	Creativity (<i>kreatif</i>)	Producing original ideas, products, or acts.

Source: (translated from <https://cerdasberkarakter.kemdikbud.go.id/profil-pelajar-pancasila/>)

Hence, in this study, a web search was conducted employing the key elements as the keywords, in order to find suitable reading texts. Besides the topic, the length of the text was also limited to between 450-900 words. The genre has to be a fictional story, in line with the course's requirement. With these criteria, the web search yielded four texts, as shown in Table 2.

Table 2*The Selected Texts, The Corresponding Pancasila Components, and The Key Elements*

No.	Title	Pancasila Components	Key Elements
1	Fear is Soft and Smooth	Autonomy	Self-regulation
2	Damon and Pythias	Nobility of Character	Inter-personal morality
3	The Beauty of Difference	Global Diversity	Intercultural communication
4	The Stone Soup Effect	Solidarity	Collaboration and teamwork

Upon finding the texts, the readability of the texts was analyzed through the Flesch-Kincaid Formula. This was done by enabling the Readability Statistic feature of the Microsoft Word software. Under the “Option” tool of Microsoft Word, the “Proofing” option was selected, and the “Show Readability Statistics” was checked. The Flesch Reading Ease values thus obtained can then be interpreted in terms of readers’ grade level, based on the average age of students who are able to comprehend the text (Gopal et al., 2021). The result is summarized here in Table 3.

Table 3*Readability Analysis Result*

No.	Title	Flesch Reading Ease	Flesch-Kincaid Grade Level	Interpretation of Flesch-Kincaid Grade Level
1	Fear is Soft and Smooth	76.5	6.4	It should be easily understood by 12- to 13-year-olds
2	Damon and Pythias	89.0	4.8	It should be easily understood by 10- to 11-year-olds
3	The Beauty of Difference	77.2	7.1	It should be easily understood by 12- to 13-year-olds
4	The Stone Soup Effect	81.2	5.6	It should be easily understood by 11- to 12-year-olds.

As the Flesch Reading Ease values all fall above 70, the four texts could be considered as fairly easy for the average adult to read (Acar & İşıSağ, 2017). Furthermore, the Flesch-Kincaid Grade Level indicated that all the texts are between 4.8 and 7.1, which is indicative of their suitability for Basic readers. Hence, it was concluded that, overall, the texts have fairly good readability.

After ensuring the readability of the texts, the next step was to draft the test specification based on the course profile and the available text materials. The resulting test specification is presented here as Appendix A. It was decided that the test format would be multiple choice, totalling 40 questions.

The drafting of the test specification was followed by the test blueprint. Each of the course’s learning outcomes and its respective indicators was mapped against Brown and Lee (2015) micro skills of reading. In turn, each micro skill was assigned to the cognitive category of Bloom’s taxonomy, in accordance with the operational verb present in the indicators. It was revealed that this Reading Fiction course stipulates that half of the indicators (three out of six) fall within the HOTS category, as can be seen in Appendix B. Lastly, upon determining that each reading text will have 10 questions, the number of questions for each indicator is specified. We also ensured that the questions are aligned with the 12 indicators of good reading comprehension performance by Grabe and Jiang (2013). The resulting test blueprint is also shown in Appendix B. Upon finalizing the blueprint, the test questions were drafted. A sample of the text (Fear is Soft and Smooth), the corresponding questions, and the answer keys is shown in Appendix C. To facilitate the data analysis, the texts and the questions were transcribed into Google Form format, and a link was generated for easy access by the participants.

Finally, when the test draft was ready, it was arranged with the instructor of the Reading Fiction course to conduct the try-out of the test with the students. It was then decided that the try-out would be carried out during the 9th meeting, two weeks after the mid-term, in mid-April 2022. One of the authors was present in the classroom to introduce the research to the students. She also explained the ethical considerations pertinent to the research, such as the voluntary nature of their participation, as well as the assurance of confidentiality. The students were also informed that their participation in the

research would not have any bearing on their final Reading Fiction grades. A written explanation of the research and the ethical aspects was also stated in the introductory part of the Google Form test. Then, after all the students showed no objection, the Google Form link was given, and the students worked on the test. Although it was planned to allow 100 minutes for the students to finish the test, it was observed that the fastest student could finish it in 26 minutes, and the last in 54 minutes. The highest score obtained by one of the participants was 80, and the lowest was 37.5. The average score was 60.2, and the median was 55.

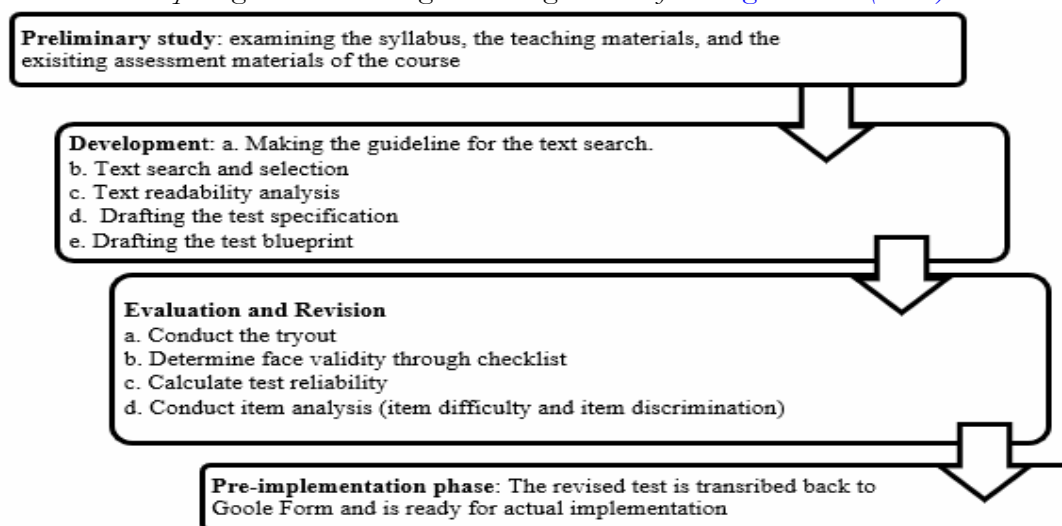
Data analysis – the evaluation and revision stages

The third part of the R&D phase is the Evaluation-Revision stage. Upon obtaining the data from the try-out, the students' responses were analysed for the test's reliability and item analysis. The reliability index was calculated by determining the Cronbach Alpha's value using SPSS 26. For the item analysis, the students' answers were checked for item difficulty and item discrimination. This was done using the ITEMAN 4.3 software. In addition, the texts and the questions were checked for face validity by providing a checklist to the Reading Fiction instructor to obtain his opinion about the suitability of the test for his students. The checklist is shown here as [Appendix D](#).

Finally, when the items which do not fulfil the item analysis criteria have been revised, the final test draft is ready for the final, pre-implementation phase. A flowchart depicting the entire design process is given in [Figure 2](#).

Figure 2

The Flowchart Depicting the R&D Design Following the Modified Borg and Gall (1983)



Determining the face validity and reliability

Although the researchers have discussed and agreed on the face validity of the test, the face validity was further ascertained by asking the opinion of the Reading Fiction lecturer, who filled out the checklist as per [Appendix D](#). He indicated that the text's length and readability were already appropriate for the students. In particular, he liked the topics of the texts, and he had the impression that the students liked them too. While he approved of the varying levels of difficulty that the questions presented, he was of the opinion that some questions were too difficult for his students. He also found some answer choices confusing, since they appear to be similar. He reported that, in his opinion, questions no. 2, 7, 10, and 36 are unclear in terms of the question formulation and the choices. This observation was noted and considered for the test improvement.

Upon completing the test, the lecturer also took the initiative to ask the students to give a rating of their understanding of the text, and also how easy they found the test. He obtained a value of 8.875 and 8.625, respectively. Hence, this constitutes another form of face validity, in which the students

themselves perceived that they understood the texts and found them manageable. However, their rather low scores on average show that they perhaps do not understand some of the details in the texts.

For reliability, the responses from the participants were tabulated and transferred into the data window of IBM SPSS 26. The reliability index, as measured by the Cronbach's alpha coefficient, was then computed. It yielded a value of 0.814, which can be categorized as "Good" (George & Mallery, 2020). The reliability analysis of ITEMAN 4.3 also generated a value of 0.832, closely resembling the output of SPSS.

Determining the item difficulty and item discrimination

The ITEMAN 3.4 software was used to compute the values of the item difficulty index (p). To run the software, the students' responses were first inserted into the data matrix file. The data from Microsoft Excel must be saved as Text format (tab-delimited). The data containing the answer key were then put in the item control file column. Afterwards, the results of the summary appear as output files (Microsoft Excel and Microsoft Word) after the data were formatted based on the columns where the data ID and the item responses began. The resulting p indices were then compared against the interpretation of the p values (see Table 4). The responses closer to 1.0 are considered as "easy" to "too easy", which means that almost all students can answer the questions correctly. The items with responses between 0.3 and 0.69 mean that the level of difficulty is "moderate" since the items are still chosen by a minimum of 30% of the students. Finally, the items are "difficult" or "too difficult" if fewer than 30% of the students chose them. This means that almost all students cannot answer the questions correctly.

Table 4

Item Difficulty Index Range and Interpretation (Rahmiati & Evaliana, 2017)

Index Range	Interpretation
Greater than 0.9	Too easy
0.7 – 0.89	Easy
0.3 – 0.69	Moderate
0.1 - 0.29	Difficult
Less than 0.09	Too difficult

The result of the Item Difficulty analysis is given in Appendix E. From the result, it can be seen that the majority of the questions fall under the category of "moderate"; 19 items or 48%. 10 Questions or 25% turned out to be "easy", and 6 questions (15%) were "too easy." Lastly, 4 questions (10%) were considered "difficult", and one question (3%) was "very difficult. A summary of the result is shown in Table 5.

Table 5

Summary of the Item Difficulty Analysis Result

Difficulty category	No. of questions	Percentage	Question numbers
Too easy	6	15%	1, 11, 20, 24, 27, 33
Easy	10	25%	2, 7, 8, 14, 15, 26, 28, 29, 32, 40
Moderate	19	48%	the remaining questions
Difficult	4	10%	30, 35, 38, 39
Too difficult	1	3%	10
Total	40		

The result of the Item Difficulty index computation was corroborated by Google Form analytics. The Google Form insights revealed that the frequently missed questions, specifically those which were answered correctly by only 1-3 students, were questions number 10, 30, 35, 38, and 39. The item discrimination index (D) was likewise obtained from ITEMAN 3.4. However, the discrimination index is represented by the software as Rpbis or the point biserial statistic index. Just

like item difficulty, each Rpbis value was categorized according to the interpretation table (see Table 6). The complete result of the Item Discrimination analysis is also given in Appendix E, while its summary is tabulated herewith as Table 7.

Table 6

Item Discrimination Index Range and Interpretation (Crocker & Algina, 2008)

Index Range	Interpretation	Action
≥ 0.40	Very good	Satisfactory
$0.30 - 0.39$	Good	Little or no revision
$0.20 - 0.29$	Marginal	Need revision
< 0.20	Poor	Eliminated or completely revised
0	No discrimination	Eliminated or completely revised
Negative value	Negative discrimination	Eliminated or completely revised

Table 7

The Summary of the Discrimination Index Analysis Results

Discrimination category	No. of questions	Percentage	Question numbers
Very good (satisfactory)	15	38%	the remaining questions
Good (little or no revision)	7	18%	2, 10, 16, 17, 21, 23, 34
Marginal (need revision)	4	10%	4, 5, 22, 36
Poor (eliminated or revised)	7	18%	7, 9, 14, 15, 26, 27, 35
Zero (eliminated or revised)	1	3%	1
Negative (eliminated or revised)	6	15%	11, 19, 20, 30, 32, 40
Total	40		

Table 7 shows that the majority of the questions, 15 in total, which represents 38%, can be considered very good in discriminating the students. Seven questions (18%) are “good” and four (10%) are “marginal.” However, quite a number (7 questions, 18%) turned out to be “poor.” One question even yielded zero value, which indicated that all students answered that question correctly. Finally, six questions (15%) produced negative Rpbis (point biserial correlation coefficient) values, which signified that they reversely discriminated against the students; in other words, the distractor (the incorrect choices) attracted the higher-ability students.

Final revision and pre-implementation

After considering all the results from the validity, reliability, and item analysis above, it was decided that some questions and their answer choices would need to be revised. After some discussions, it was decided to revise the questions in a manner tabulated in Table 8, which also shows the rationale behind such revision. In sum, questions 1, 11, 20, and 27 were almost completely revised, while questions 19, 30, 32, and 40 were revised only for those options (A, B, C, or D) which were not chosen at all by the students, which indicated that they were not effective as distractors. The final, revised questions are available at <https://bit.ly/PancasilaReadingTest>.

Table 8

The Revised Questions, the Reasons, and the Actions Taken

Revised Question Number	Reason	Actions
1	Too easy and not discriminating (zero value)	Revise the question and/or the choices
11	Too easy and negatively discriminating	
20	Too easy and negatively discriminating	
27	Too easy and poorly discriminating	
19, 30, 32, 40	Negative discrimination	Revise distractors
7	Poor face validity and poorly discriminating	Revise the question formulation

DISCUSSION

This study sets out to model the development of a valid, reliable, and original reading comprehension test that integrates higher order thinking skills (HOTS) with character education, framed by the Pancasila Student Profile. The originality of this test lies in its dual innovation: first, it systematically maps reading sub-skills onto HOTS domains to ensure that students engage in higher-level cognitive processes; second, it embeds multiple Pancasila values into narrative texts and comprehension questions, thereby linking cognitive development with moral and civic formation. While [Rahmiati and Emaliana \(2017\)](#) focused on HOTS-based reading comprehension and [Sudarwati et al. \(2021\)](#) incorporated only honesty into grammar items, our study uniquely combines HOTS with a balanced representation of Pancasila values. This approach makes the instrument distinct from existing assessments, as it simultaneously advances evaluative thinking and character-building in line with the Merdeka Curriculum.

This study also responds to and validates the findings of [Ratrisari et al. \(2024\)](#), who investigated the perceptions of Indonesian Language teachers in grade VII on the essential aspects in constructing a summative assessment guideline with HOTS and Pancasila Student Profile integration. In their study, the teachers proposed that HOTS and Pancasila Student Profile be incorporated under the reading-viewing skill, using descriptive, folk poetry, and narrative texts as the more suitable genre. Our study is then aligned with theirs by choosing a narrative text as the medium of infusing HOTS and Pancasila Student Profile values, albeit in a differing context of English language assessment in higher education. Next, [Ratrisari et al. \(2024\)](#)'s study also illustrated the teachers' opinions on the proportion of Pancasila Student Profile's values that should be included in a narrative text, with the highest given to Global Diversity, and lowest to Critical Reasoning. In this sense, our study provides an alternative view from theirs by presenting four narrative texts, each emphasizing four values from the Pancasila Student Profile with an equal number of questions, thus giving a more balanced outlook of the profile.

However, [Ratrisari et al. \(2024\)](#)'s study also illustrated the teachers' aspiration as to the ratio of HOTS-loaded questions, namely 20:60:20 for C4, C5, and C6, respectively. In other words, the greatest percentage should be given to C5 (evaluating) type of questions, and lesser to C4 (analyzing) and C6 (creating). Our study thus diverges from this proposal by including questions in the C2 (understanding) and C3 (applying) levels and not having any questions at the C6 level. Nonetheless, C5 constituted 50% of the questions. In this sense, we believe that a certain proportion of the questions should still contain LOTS, while acknowledging that crafting a C6-level question remains challenging. Within the Pancasila framework, C5 aligns with values of autonomy and critical thinking, equipping learners to make ethical and informed judgments. Thus, giving the greatest percentage to C5 questions is justified both pedagogically and ideologically. Our study balanced HOTS with selected LOTS items, acknowledging the need for scaffolding and the difficulty of crafting valid C6 questions. Nonetheless, by allocating 50% of items to C5, we align with their emphasis on evaluation while adapting to practical constraints.

The second research question concerned the measures taken to ensure that the reading assessment is valid and reliable. This study employed four types of analysis, namely face validity, internal reliability, item difficulty, and item discrimination. In this study, the face validity was obtained by providing a checklist for the instructor of the course to fill out, so that his opinion of the reading texts and the questions can be sought, since he knows the students well. Other studies carried out expert validity ([Özdemir & Akyol, 2019](#); [Rahmiati & Emaliana, 2017](#); [Setiawan et al., 2019](#)), or content and construct validity ([Hanafi, 2016](#)). [Özdemir and Akyol \(2019\)](#) even asked two students to read the text and the questions before the try out, to check for the test's understandability. Subsequent to the test, [Sudarwati et al. \(2021\)](#) conducted a survey to the students to garner their perspective on how the test had helped them and their general impression of the test. On the other hand, [Ningrum and Sudarwati \(2022\)](#) conducted a peer debriefing by asking four validators to give feedback to the test blueprint. In this study, the content and construct validity were presumably valid, since the test was developed in a way tailored to the course type and requirements, as depicted by the test specification ([Khairani & Shamsuddin, 2016](#)). This study deemed that face validity was sufficient for the research

purpose without obtaining expert validation. Asking students to read the texts beforehand was also regarded as unnecessary, due to the optimal values of the texts' readability index, as shown in [Table 3](#).

Upon analysing the test items, those which were deemed unsatisfactory in terms of item difficulty and discrimination were carefully revised to improve their quality. This revision process followed the practice of [Hanafi \(2016\)](#) and [Rahmiati and Emaliana \(2017\)](#), who emphasized that modifying items can enhance their alignment with learning objectives while maintaining overall test validity. In contrast, other studies, such as [Özdemir and Akyol \(2019\)](#) and [Setiawan et al. \(2019\)](#) opted to discard problematic items altogether, prioritizing efficiency over refinement. Our decision to revise rather than eliminate reflects a commitment to preserving the integrity of the test blueprint and ensuring that each reading sub-skill, HOTS domain, and Pancasila value remains represented. This approach also resonates with [Ware & Vik \(2009\)](#), who argue that item quality can often be improved through adjustment of distractors and wording, thereby strengthening discriminatory power. Thus, the revisions made in this study not only improved reliability but also safeguarded the originality of the instrument by retaining its unique integration of HOTS and character education.

In the study of [Khairani and Shamsuddin \(2016\)](#) it was revealed that, on average, only 67% of the test questions satisfied the criteria of having acceptable item difficulty and discrimination index at the same time ($0.26 \leq p \leq 0.75$ and $D \geq 0.20$). Although the present study did not employ such stringent criteria as those of [Khairani and Shamsuddin \(2016\)](#), it can be approximated that the 9 revised questions constituted 77.5% of the total items. Overall, this finding shows that it remains a challenge to develop test questions that satisfy both indices simultaneously. While item difficulty can be constructed with Bloom's taxonomy as the guide, [Khairani and Shamsuddin \(2016\)](#) lamented that there is currently no clear guideline on how to develop test items with acceptable discriminatory power. They further indicated that the key to success lies in the distractors, which have to be selected by at least 5% of the test takers ([Ware & Vik, 2009](#)). Finally, they suggested that teachers develop their own item banks, which contain test items with the item analysis parameters stated.

Lastly, having a valid and reliable assessment instrument does not guarantee that the test is suitable for or meets the needs of the students. As shown by the findings of [Ningrum and Sudarwati \(2022\)](#), despite the test items having difficulty levels ranging between easy and moderate, their students still scored rather low in the test piloting. Our participants also obtained an average score of 60, which was not very satisfactory either, although the majority of our questions fell under the moderate range. [Ningrum and Sudarwati \(2022\)](#) conjectured that it might be due to the few try-out participants ($n = 20$) or the low critical reading proficiency of the students. We concur with them that our result may not be generalizable due to the small number of participants ($n = 12$). In addition, other contextual factors might have come into play, such as the students' earnestness in doing the test, knowing that it was done for research and would not go into their final grades. Nonetheless, we believe that this research is valuable in detailing the development of a valid and reliable reading assessment instrument infused with HOTS and Pancasila Student Profile values.

Interpreted through the lens of language assessment theory, the development of this instrument demonstrates how validity and reliability can be achieved while also embedding broader educational goals. The systematic mapping of reading sub-skills onto HOTS domains reflects the pedagogical principle that assessment should drive higher-order learning rather than reinforce rote comprehension. At the same time, the integration of Pancasila values into reading texts and questions illustrates how assessment can serve as a vehicle for character education, aligning with perspectives that view testing not only as a measure of learning but also as a formative tool for shaping dispositions and civic responsibility ([Lickona et al., 2002](#)). By combining HOTS pedagogy with character education integration, this study advances a holistic model of classroom assessment that supports both cognitive and moral development, thereby responding to the dual mandate of the Merdeka Curriculum.

The present study carries some practical implications. For teachers, the model demonstrates how HOTS-based assessments can be designed within classroom constraints. While time and resources remain challenges, the blueprint offers a replicable framework. Teachers can scaffold

evaluative questions by gradually moving from LOTS to HOTS, ensuring accessibility for diverse student abilities. This practical implication is crucial for bridging policy mandates with classroom realities. Overall, this study contributes a validated model for HOTS-based reading assessment that prioritizes evaluative thinking while embedding character education through Pancasila values. By explaining mechanisms, contrasting with related works, and acknowledging limitations, the study offers both theoretical insights and practical guidance. It demonstrates that classroom-based assessments can simultaneously foster higher-order thinking and character development, thereby advancing both academic and civic goals.

This study acknowledges limitations, including the small sample size ($n=12$) and the single-course context, which restrict generalizability. Future research should expand into larger cohorts, cross-institutional settings, and digital platforms. Moreover, due to time constraints, this study also chose to apply less stringent criteria in revising the test items; not all questions with $D < 0.20$ were revised. Thus, the test items could have been perfected by adhering to the conventional discrimination criteria. For future research, immediate and short-term follow-up studies can be conducted to implement and analyse the revised assessment (<https://bit.ly/PancasilaReadingTest>) to determine whether there are improvements in the item analysis results. Besides, the students' opinions can also be obtained through a survey or questionnaire to unearth their perspective on the various aspects, cognitive and practical, of the test. In addition, this research can also be continued by analysing the relationship between the cognitive level of the questions (as indicated by the HOTS' mapping of the questions) and the item difficulty, to find whether the HOTS questions are truly challenging for the students.

CONCLUSION

This study has modelled the development and validation of an original reading comprehension test integrating HOTS and character education, specifically Pancasila values. Each step of the development was described in accordance with the R&D design model. HOTS were utilized in the test specification in order to map the cognitive level of the question types. Pancasila value indicators were employed to select the reading texts. The resulting 4 texts and 40 multiple-choice questions were tried out on 12 undergraduate students. The result was analysed for face validity, internal reliability, item difficulty, and item discrimination. The internal reliability was found to be good (0.812), while feedback from face validity and the output from the item analysis were utilized to revise a total of 9 questions. The products of this study, namely a description of the procedure in developing an original assessment as well as the assessment itself, are expected to assist EFL teachers in their quest to make valid and reliable test items. The findings can also be useful for teacher training institutions in developing teaching materials for assessment courses. The result of this study also shows that developing an original assessment is certainly not an easy task, but not impossible either. With practice, teachers can get more experience and confidence in test development. Nevertheless, this study reiterates the call from the previously cited literature for more preservice teacher training or in-service professional development programs to guide teachers in developing authentic, valid, and reliable tests. Beyond these immediate contributions, the study also offers broader implications: for ELT teachers, it provides a replicable framework for designing assessments that balance cognitive rigor with character formation; for curriculum developers, it demonstrates how assessment instruments can be aligned with the Merdeka Curriculum's dual mandate of fostering evaluative thinking and civic responsibility; and for teacher education programs, it underscores the importance of embedding assessment literacy into training so that future teachers are equipped to design instruments that are both pedagogically sound and values-oriented.

ACKNOWLEDGEMENTS

We would like to thank the Head of Department and the faculty member of the English Language Education Study Program, Teacher Education Faculty, Universitas Katolik Widya Mandala Surabaya for their assistance and cooperation in this study.

REFERENCES

- Abdulkarim, A., Komalasari, K., Saripudin, D., Ratmaningsih, N., & Anggraini, D. N. (2020). Development of a unity in diversity-based Pancasila education text book for Indonesian universities. *International Journal of Instruction*, 13(1), 371–386. <https://doi.org/10.29333/iji.2020.13125a>
- Acar, A., & İŞİSağ, K. U. (2017). Readability and comprehensibility in translation using reading ease and grade indices. *International Journal of Comparative Literature and Translation Studies*, 5(2), 47. <https://doi.org/10.7575/aiac.ijclts.v.5n.2p.47>
- Afflerbach, P. (2016). Reading assessment: Looking ahead. *The Reading Teacher*, 69(4), 413–419. <https://doi.org/10.1002/trtr.1430>
- Ahmadi, F., Rochmad, R., Lestari, F. P., & Harjunowibowo, D. (2021). The development of mathematics comic containing Pancasila values to develop character of elementary school students: A case study of Indonesia. *Journal of Innovation in Educational and Cultural Research*, 2(1), 25–34. <https://doi.org/10.46843/jiecr.v2i1.20>
- Amali, L. N., Anggani Linggar Bharati, D., & Rozi, F. (2022). The implementation of high order thinking skills (HOTS) assessment to evaluate the students' reading comprehension achievement. *English Education Journal*, 12(1), 10–18. <https://doi.org/10.15294/eej.v12i1.52571>
- Ardiana, A. P., & As Sabiq, A. H. (2020). HOTS-based analysis on English reading comprehension formative assessment. *Tarling: Journal of Language Education*, 4(1), 19–36. <https://doi.org/10.24090/tarling.v4i1.3975>
- Arifin, I., Wilujeng, I., & Jumadi, J. (2020). The effect of quick on the draw model assisted by the physics learning book integrated Pancasila values on critical thinking skill. *Jurnal Penelitian & Pengembangan Pendidikan Fisika*, 6(1), 121–130. <https://doi.org/10.21009/1.06114>
- Arshad, A. (2020). A review on the principles of a reading comprehension test construction to assess the test takers at different levels. *Psychology and Education*, 57(8), 1290-1302.
- Atjonen, P. (2014). Teachers' views of their assessment practice. *The Curriculum Journal*, 25(2), 238–259. <https://doi.org/10.1080/09585176.2013.874952>
- Atjonen, P., Pöntinen, S., Kontkanen, S., & Ruotsalainen, P. (2022). In enhancing preservice teachers' assessment literacy: Focus on knowledge base, conceptions of assessment, and teacher learning. *Frontiers in Education*, 7, 891391. <https://doi.org/10.3389/educ.2022.891391>
- Bachman, L. F. (2010). *Fundamental considerations in language testing* (1. publ., 11. [print.]). Oxford University Press.
- Berkowitz, M. W. (2012). *Understanding Effective Character Education*. CSEE Connections, the Center for Spiritual and Ethical Education. <https://www.researchgate.net/publication/265068585>
- Borg, W. R., & Gall, M. D. (1983). *Educational research: An introduction*. Longman
- Bower, M., Torrington, J., Lai, J. W. M., Petocz, P., & Alfano, M. (2024). How should we change teaching and assessment in response to increasingly powerful generative Artificial Intelligence? Outcomes of the ChatGPT teacher survey. *Education and Information Technologies*. <https://doi.org/10.1007/s10639-023-12405-0>
- Brata, D. P. N., Utomo, E. S., & Sukardi, S. (2022). *The analysis of students' attitudes construction based on Pancasila profile to be integrated with teacher's lesson plan in junior high school in pandemic era*. 2nd International Conference on Education and Technology (ICETECH 2021). <https://doi.org/10.2991/assehr.k.220103.045>
- Brown, H. D., & Lee, H. (2015). *Teaching by principles: An interactive approach to language pedagogy*. Pearson Education. https://books.google.co.id/books/about/Teaching_by_Principles.html?hl=id&id=4XEbrgEACAAJ&redir_esc=y
- Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition*, 7(3), 323–331. <https://doi.org/10.1016/j.jarmac.2018.07.002>
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning

- DeLuca, C., & Bellara, A. (2013). The current state of assessment education: Aligning policy, standards, and teacher education curriculum. *Journal of Teacher Education*, 64(4), 356–372. <https://doi.org/10.1177/0022487113488144>
- Faradella, N. E., Wiyaka, W., & Egar, N. (2024). Teachers' strategies and challenges in conducting HOTS-based activities in Merdeka Curriculum era. *Indonesian Journal of Education and Pedagogy*, 1(2), 75–90. <https://doi.org/10.61251/ijoep.v1i2.64>
- George, D., & Mallery, P. (2020). *IBM SPSS statistics 26 step by step*. Routledge
- Gopal, R., Maniam, M., Madzlan, N. A., Shukor, S. S. B., & Neelamegam, K. (2021). Readability formulas: An analysis into reading index of prose forms. *Studies in English Language and Education*, 8(3), 972–985. <https://doi.org/10.24815/siele.v8i3.20373>
- Grabe, W., & Jiang, X. (2013). Assessing reading. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed., pp. 185–200). Wiley. <https://doi.org/10.1002/9781118411360.wbcla060>
- Gustiani, S. (2019). Research and development (R&D) method as a model design in educational research and its alternatives. *Holistics (Hospitality and Linguistics): Jurnal Ilmiah Bahasa Inggris*, 11(2). 11(2).
- Hanafi, H. (2016). Developing reading comprehension test for the first semester students of english department. *ELLITE: Journal of English Language, Literature, and Teaching*, 1(1). <https://doi.org/10.32528/ellite.v1i1.164>
- Haniah, A. R., Aman, A., & Setiawan, R. (2020). Integration of strengthening of character education and higher order thinking skills in history learning. *Journal of Education and Learning (EduLearn)*, 14(2), 183–190. <https://doi.org/10.11591/edulearn.v14i2.15010>
- Hidayat, M., Rozak, R. W. A., Hakam, K. A., Kembara, M. D., & Parhan, M. (2021). Character education in Indonesia: How is it internalized and implemented in virtual learning? *Jurnal Cakrawala Pendidikan*, 41(1), 186–198. <https://doi.org/10.21831/cp.v41i1.45920>
- Hoang, L. P., Le, P. A., Le, H. T., Nguyen, D. T., Phan, T. C., Le, H. T., Van Tran, H., & Tran-Duong, Q. H. (2025). Evaluating educational assessment competence of pre-service teachers: Extended standards in the context of digital classroom assessment transformation. *Education and Information Technologies*, 30(12), 16347–16374. <https://doi.org/10.1007/s10639-025-13467-y>
- Khairani, A. Z., & Shamsuddin, H. (2016). Assessing item difficulty and discrimination indices of teacher-developed multiple-choice tests. In S. F. Tang & L. Logonnathan (Eds.), *Assessment for learning within and beyond the classroom* (pp. 417–426). Springer Singapore. https://doi.org/10.1007/978-981-10-0908-2_35
- Khorsand, N. (2009). *Cognitive levels of questions used by Iranian EFL teachers in advanced reading comprehension tests*. <https://www.semanticscholar.org/paper/Cognitive-Levels-of-Questions-Used-by-Iranian-EFL-Khorsand/e173875b81e2036cfc68b5c569dbea07a2ceb523>
- Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2
- Lickona, T., Schaps, E., & Lewis, C. (2002). Eleven principles of effective character education. *Special Topics, General*. <https://digitalcommons.unomaha.edu/slcestdgen/50>
- Magliano, J. P., Millis, K., Ozuru, Y., & McNamara, D. S. (2007). A multidimensional framework to evaluate reading assessment tools. In *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 107–136). Lawrence Erlbaum Associates Publishers.
- Miller, M. D., Linn, R. L., Gronlund, N. E., & Linn, R. L. (2009). *Measurement and assessment in teaching* (10th ed). Merrill/Pearson.
- Nenohai, J. (2021). Internalizing character education (CE) into English language teaching: Teachers' perspectives. *Jurnal Manajemen Pendidikan*, 8(2).
- Newton, P. M., & Essex, K. (2024). How common is cheating in online exams and did it increase during the COVID-19 pandemic? A systematic review. *Journal of Academic Ethics*, 22(2), 323–343. <https://doi.org/10.1007/s10805-023-09485-5>

- Ningrum, A. S. B., & Sudarwati, E. (2022). Is the test sensible? Developing a critical reading test for Indonesian EFL learners. *JEELS (Journal of English Education and Linguistics Studies)*, 9(1), 209–227. <https://doi.org/10.30762/jeels.v9i1.4005>
- Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessment of students*. Pearson Education.
- Oktarina, Y., Inderawati, R., & Petrus, I. (2022). Developing local culture-based EFL reading materials for the 21st-century learning. *Studies in English Language and Education*, 9(3), 1128–1147. <https://doi.org/10.24815/siele.v9i3.24660>
- Oktriani, D. R., Hufad, A., & Utami, N. F. (2023). Overcoming the character crisis in children: strategies, outcomes, and evaluations of Bina Desa program. *Utamax: Journal of Ultimate Research and Trends in Education*, 5(2). <https://doi.org/10.31849/utamax.v5i2.13758>
- Pastore, S. (2023). Teacher assessment literacy: A systematic review. *Frontiers in Education*, 8, 1217167. <https://doi.org/10.3389/feduc.2023.1217167>
- Putra, T. K., & Abdullah, D. F. (2019). Higher-order thinking skill (HOTS) questions in English national examination in Indonesia. *Jurnal Bahasa Lingua Scientia*, 11(1), 145–160. <https://doi.org/10.21274/lis.2019.11.1.145-160>
- Rachmawati, D. L., & Purwati, O. (2021). EFL teachers' attitudes and competence in developing HOTS-based formative assessment. *JEES (Journal of English Educators Society)*, 6(2). <https://doi.org/10.21070/jees.v6i2.1060>
- Rahmiati, I. I., & Emaliana, I. (2017). Developing reading test using lower to higher order of thinking for ESP students. *Language in India*, 17(11), 124–144.
- Ratrisari, A. A., Subyantoro, & Wagiran, W. (2024). Teachers' perceptions of the development of Indonesian language assessment guidelines containing HOTS and Pancasila student profiles. *Seloka: Jurnal Pendidikan Bahasa Dan Sastra Indonesia*, 13(1), 27–36.
- Rofii, A., Murtadho, F., & Rahmat, A. (2018). Model of contextual-based academic writing learning module. *English Review: Journal of English Education*, 6(2), 51. <https://doi.org/10.25134/erjee.v6i2.1242>
- Sackett, P. R. (2012). Cognitive tests, constructs, and content validity: A commentary on Schmidt. *International Journal of Selection and Assessment*, 20(1), 24–27. <https://doi.org/10.1111/j.1468-2389.2012.00576.x>
- Sari, A. M., & Ariswan, A. (2021). The integrated physics learning e-module with Pancasila character values in work and energy subjects as solution to improve students' critical thinking ability and independence: Is it effective? *Jurnal Ilmiah Pendidikan Fisika Al-Biruni*, 10(1), 85. <https://doi.org/10.24042/jipfalbiruni.v10i1.7749>
- Sari, A. M., Wilujeng, I., & Satriana, A. (2020). The development of optical devices learning book integrated with Pancasila practice values. *Journal of Physics: Conference Series*, 1440(1), 012036. <https://doi.org/10.1088/1742-6596/1440/1/012036>
- Sato, T., & Ikeda, N. (2015). Test-taker perception of what test items measure: A potential impact of face validity on student learning. *Language Testing in Asia*, 5(1), 10. <https://doi.org/10.1186/s40468-015-0019-z>
- Seifert, S., & Paleczek, L. (2021). Digitally assessing text comprehension in grades 3-4: Test development and validation. *Electronic Journal of E-Learning*, 19(5), 336–348. <https://doi.org/10.34190/ejel.19.5.2467>
- Setiawan, H., Khair, B. N., Ratnadi, R., Hakim, M., & Istiningsih, S. (2019). Developing HOTS-based assessment instrument for primary schools. *Proceedings of the 1st Annual Conference on Education and Social Sciences (ACCESS 2019)*. 1st Annual Conference on Education and Social Sciences (ACCESS 2019). <https://doi.org/10.2991/assehr.k.200827.054>
- Shofa, N. A. (2021). *Strengthening the profile of Pancasila students through project-based learning in learning history*. In *Prosiding Seminar Nasional & Internasional*.
- Sudarwati, E., Fatimah, F., Astuti, Y., & Ubaidillah, M. F. (2021). Developing online learning assessment instrument for English sentence structure course during Covid-19 pandemic.

- Langkawi: Journal of The Association for Arabic and English*, 7(2), 170. <https://doi.org/10.31332/lkw.v7i2.3122>
- Tyas, M. A., Nurkamto, J., Marmanto, S., & Laksani, H. (2019). *Developing higher order thinking skills (HOTS)-Based Questions: Indonesian EFL Teachers' Challenges*. 52–63. <https://doi.org/10.17501/26307413.2019.2106>
- Ware, J., & Vik, T. (2009). Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. *Medical Teacher*, 31(3), 238–243. <https://doi.org/10.1080/01421590802155597>
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach. hampshire*. Palgrave McMillan, 10, 9780230514577.
- Wijayanti, P. S., Jamilah, F., Herawati, T. R., & Kusumaningrum, R. N. (2022). Penguatan penyusunan modul proyek profil pelajar Pancasila pada sekolah penggerak jenjang SMA [*Strengthening the development of Pancasila student profile project modules in activator schools at the senior high school level*]. *ABDIMAS NUSANTARA: Jurnal Pengabdian Kepada Masyarakat*, 3(2), 43–49.
- Wuryandani, W., Sartono, E. K. E., & Fathurrohman. (2020). Development of Big Book media to increase understanding of Pancasila values in early class. *Proceedings of the International Conference on Educational Research and Innovation (ICERI 2019)*. International Conference on Educational Research and Innovation (ICERI 2019). <https://doi.org/10.2991/assehr.k.200204.023>
- Zakso, A., Agung, I., Sofyatinigrum, E., & Capnary, M. C. (2022). *Factors affecting character education in the development of the profile of Pancasila students: The case of Indonesia*. 6(2), 2254–2273.
- Özdemir, E. Ç., & Akyol, H. (2019). The development of a reading comprehension test. *Universal Journal of Educational Research*, 7(2), 563–570. <https://doi.org/10.13189/ujer.2019.070229>

CONFLICT OF INTEREST STATEMENT:

The authors affirm no conflicting interests regarding the research in this manuscript.

About the authors

Imelda Gozali is a faculty member of the English Language Education Study Program, Teacher Education Faculty, Universitas Katolik Widya Mandala Surabaya. She obtained her doctoral degree from the Universitas Negeri Malang. Her research area is in the field of assessment and feedback, as well as the use of technology for language teaching. ORCID: <https://orcid.org/0000-0003-1276-6452>, email: imelda.gozali@ukwms.ac.id

Faizatul Istiqomah is a lecturer at the Faculty of Education and Teacher Training of Universitas Islam Negeri Sayyid Ali Rahmatullah Tulungagung, Indonesia. She earned her Master degree in Education of TESOL International from Monash University of Melbourne, Australia. Her research interests include English language teaching, teaching grammar and academic writing, and systemic functional linguistics. ORCID: <https://orcid.org/0000-0002-5288-3760>, email: faizatulistiwa79@gmail.com.

Ly Gia Ngan Vo is currently pursuing a Master's degree in TESOL at the University of Social Sciences and Humanities, Vietnam National University Ho Chi Minh City. She holds a Bachelor's degree in English Linguistics with a specialization in English Language Teaching and is also TESOL-certified. In addition to her graduate studies, Ngan serves as a visiting lecturer at the University of Social Sciences and Humanities, where she teaches English to non-English major students. She has extensive experience teaching English proficiency and communication courses, including IELTS and TOEIC preparation, and has worked with learners ranging from teenagers to university students. Throughout her academic and professional career, she has developed expertise in curriculum design, speaking assessment, and instructional support for mixed-ability learners. Her research interests include linguistic skills development, assessment, and the integration of technology into language instruction. ORCID: <https://orcid.org/0009-0008-3933-9921>, email: volygiangan@hcmussh.edu.vn

Appendix A

Test Specification

No	Test Specification	
1	Course Name:	Reading Fiction
2	Level:	Undergraduate (English Department, Faculty of Teacher Training and Education)
3	Semester:	II
4	Credit hours:	2
5	Course Description:	The course provides exercises in meaning reconstruction of intermediate level reading passages. It trains the students in identifying specific and supporting details, inferring main ideas, and creating evaluative comments about the contents of the texts
6	[The English Department] Expected Learning Outcome (ELO):	Graduates are able to communicate various topics to various audiences in spoken and written English in various contexts, in various genres, and for various purposes effectively
7	Course Learning Outcome (CLO):	CLO 1 Students are able to comprehend the context of the reading. In particular, students are able to: CLO 1.1 Predict the topic of the discourse/text. CLO 1.2 Identify important pieces of information. CLO 1.3 Summarize the overall content of the text CLO 2 Students are able to connect the reading to their daily life. In particular, students are able to: CLO 2.1 Evaluate ideas presented in the texts. CLO 2.2 Predict probable outcomes of certain ideas CLO 2.3 Predict the application of certain ideas to real-life situations.
8	Test Type:	Multiple choice
9	Text Type:	4 (four) Narrative Texts 1. Fear is soft and smooth (For Autonomy) 2. Damon and Pythias (For Nobility of Character) 3. The Beauty of Difference (For Global Diversity) 4. The Stone Soup Effect (For Solidarity)
10.	Sources of Texts:	1. https://freestoriesforkids.com/children/stories-and-foes/fear-soft-and-smooth 2. https://americanliterature.com/author/james-baldwin/short-story/damon-and-pythias 3. https://www.kidsout.org.uk/world-stories-wednesday-beauty-difference/#:~:text=This%20week's%20story%20is%20called,it's%20good%20to%20be%20different 4. https://purposefocuscommitment.com/teamwork-story-stone-soup-effect/
11.	Total no. of items:	40
12.	Time allocation:	100 minutes
13.	Materials needed:	Paper and pencil
14.	Scoring:	1 for correct answers, 0 for incorrect answers The formula for total score is: $\text{Total Score} = \frac{\text{Score obtained}}{\text{Maximum score}} \times 100\%$

Appendix B*Test Blueprint*

Assessment Goals based on Course Learning Outcome (CLO) 1 and 2	Indicators based on the CLOs	Reading Micro Skills (Brown & Lee, 2015)	Bloom Classification (HOTS)	Number of items	Percentage
Given four (4) reading texts and within 100 minutes, students are able to comprehend the context of the reading and connect the reading to their daily lives	CLO 1.1. Students are able to predict the topic of the discourse/text.	Topic identification	Understanding	1	10%
	CLO 1.2 Students are able to identify important pieces of information.	(a) supporting ideas/details (b) Pronoun reference (c) Word meaning	Analyzing	(a) 1 (b) 1 (c) 1	30%
	CLO 1.3 Students are able to summarize the overall content of the text	Main idea identification and text organization	Understanding	1	10%
	CLO 2.1 Students are able to evaluate ideas presented in the texts	(a) implied details (understanding, description) (b) the writer's tone of writing	Evaluating	(a) 1 (b) 1	20%
	CLO 2.2. Students are able to predict probable outcomes of certain ideas	implied details (sequence, and cause- effect)	Analyzing	2	20%
	CLO 2.3 Students are able to predict the application of certain ideas to real-life situations.	implied details (problem solving)	Applying	1	10%
TOTAL				10	100%

Appendix C

Test Form

DIRECTIONS:

1. Write your name and student number on the answer sheet provided.
2. Read the texts and write the answers on the answer sheet provided by using capital letter (A, B, C, or D)
3. You may not use the dictionary.
4. The test consists of 4 (four) texts and 40 questions. You have 100 minutes to finish the test.
5. Do not write anything on the test book.

TEXT 1. For Questions No. 1-10

Fear is Soft and Smooth

Marina was extremely afraid of the dark. When the lights went out, everything and every shadow appeared to her as the most terrible of monsters. Her parents explained to her, every day and with great patience, that these things were not monsters. Marina understood her parents, but she could not stop feeling an awful fear whenever it was dark.

One day her Aunt Valerie came to visit. Valerie was an incredible woman. She was famous for her courage, and for having gone on many journeys of adventure, some of which had been made into books and movies. Marina wanted to conquer her fear of the dark, so she asked her aunt how she became so brave, and whether she had ever been frightened.

"A great many times, Marina," answered her aunt, "I remember when I was small and I was terribly afraid of the dark. I couldn't stay in the dark for even a moment".

Marina became very excited. How was it possible that someone so courageous could have been afraid of the dark?

"I'll tell you a secret, Marina. It was some blind children who taught me how to be brave. They can't see, so if they had never discovered the secret of how not to be afraid of the dark, they would have been forever frightened". "It's true!" said Marina, intrigued, "Can you tell me that secret?" "Of course! The secret is to change your eyes. Since blind children can't see, their hands are their eyes. All you have to do to conquer your fear is what they do. Shut the eyes of your face and open the eyes of your hands. Let's make a deal: tonight, when you go to bed and put out the light, if anything makes you afraid, close your eyes, carefully get out of bed, and try to see what it is that's making you scared. But do it using your hands as eyes... and tomorrow tell me how you're getting on with the fear".

Marina accepted, but she was rather worried. She knew she would need to be brave to close her eyes and go and touch whatever it was that was frightening her; but she was willing to try because she was already too old for this. When her parents took her to bed, she herself put out the light. After a little while, she felt afraid of one of the shadows in the bedroom. Following the advice of Aunt Valerie, she closed the eyes of her face and opened the eyes of her hands. And, summoning up all her courage, she went over to touch that mysterious shadow..

The next morning Marina came running into the kitchen, a big smile on her face, and a song on her lips. "The monster is so soft and smooth!" she cried, "It's my teddy bear!"

Source: <https://freestoriesforkids.com/children/stories-and-tales/fear-soft-and-smooth>

1. What is the text about? (topic)
 - a. Fear of the Monster
 - b. Fear of the dark
 - c. Fear of sleeping alone
 - d. Fear of teddy bear
2. How did Marina's mother and father tell her about the monsters she was afraid of? (Important piece of information)
 - a. Gently
 - b. Slowly
 - c. Occasionally
 - d. Clearly

3. Find the word “which” in bold in the second paragraph. What does “which” refer to? (Pronoun reference)
 - a. Aunt Valerie
 - b. Marina
 - c. Journeys
 - d. Books
4. This statement is true about Aunt Valerie, EXCEPT? (main idea identification)
 - a. She was also very afraid of the dark
 - b. She had traveled to numerous places
 - c. She was one of the blind children
 - d. She wrote some of her stories.
5. Why did Marina talk to Aunt Valerie? (problem solving)
 - a. Because Marina was interested to have a brave adventure
 - b. Because Marina planned to visit her house and read her book
 - c. Because Marina was curious if Aunt Valerie was also afraid of something.
 - d. Because Marina wanted to tell her story in the dark
6. What is the synonym of “courageous”? (word meaning)
 - a. Popular
 - b. Gorgeous
 - c. Adventurous
 - d. Daring
7. How did Marina follow the secret instruction from her aunt? (sequence)
 - a. By closing her eyes and covering her face in the dark.
 - b. By closing her eyes and moving her hands in the dark.
 - c. By covering her face in the dark and closing her eyes.
 - d. By moving her hands in the dark and closing her eyes.
8. What did Marina feel in the following day after trying to practice her aunt’s secret? (implied detail)
 - a. Sad
 - b. Afraid
 - c. worried
 - d. happy
9. What would probably Marina do later in the following nights when she was in the dark after knowing the secret? (cause effect)
 - a. She would close her eyes and move her hands in the dark
 - b. She would not be afraid anymore.
 - c. She would stay away from her teddy bear.
 - d. She would not be afraid of monsters.
10. After listening to the advice of Aunt Valerie, Marina has the following character except:(writer’s tone or purpose of writing)
 - a. Self-actualization
 - b. Self-regulation.
 - c. Self-control
 - d. Self-awareness

Appendix D

Test Review Form

TEST REVIEW FORM

Course Name: Reading Fictions

Course Code / Semester: EGL158/2

Credit Units: 2

No	Reviewed Points	YES **)	NO **)	COMMENT
1.	The topics of the reading texts are in accordance with the syllabus of this course			
2	The length of the texts in the test is appropriate for this course			
3	The readability (sentence length, unique words) of the texts in the test is appropriate for this course			
4	The time allocated for this test is sufficient			
5	The cognitive level of the questions is appropriate for the students			

Day/date:

Reviewer's name:

Reviewer's signature: _____

Appendix E*The Results of Item Difficulty and Item Discrimination Analysis*

The Results of Item Difficulty and Item Discrimination Analysis by ITEMAN 4.3

No. item	Item difficulty (P)	Results	Item discrimination (Rpbis)	Choice (Prop. Endorsing)				Suggestion to item
				A	B	C	D	
1	1.000	Too easy	0.000	0.000	1.000	0.000	0.000	eliminated or completely revised
2	0.750	easy	0.387	0.750	0.167	0.000	0.083	little or no revision
3	0.500	moderate	0.673	0.500	0.000	0.500	0.000	satisfactory
4	0.667	moderate	0.246	0.250	0.083	0.667	0.000	need revision
5	0.667	moderate	0.213	0.250	0.000	0.667	0.083	need revision
6	0.583	moderate	0.683	0.167	0.083	0.167	0.583	satisfactory
7	0.833	easy	0.181	0.083	0.833	0.000	0.083	eliminated or completely revised
8	0.833	easy	0.470	0.000	0.000	0.167	0.833	satisfactory
9	0.417	moderate	0.050	0.417	0.417	0.000	0.167	eliminated or completely revised
10	0.083	Too difficult	0.381	0.083	0.000	0.500	0.417	little or no revision
11	0.917	Too easy	-0.044	0.000	0.917	0.083	0.000	eliminated or completely revised
12	0.667	moderate	0.512	0.083	0.250	0.000	0.667	satisfactory
13	0.667	moderate	0.512	0.667	0.250	0.083	0.000	satisfactory
14	0.750	easy	0.034	0.083	0.750	0.083	0.083	eliminated or completely revised
15	0.750	easy	0.103	0.000	0.250	0.000	0.750	eliminated or completely revised
16	0.333	moderate	0.355	0.083	0.500	0.333	0.083	little or no revision
17	0.333	moderate	0.355	0.333	0.167	0.333	0.167	little or no revision
18	0.500	moderate	0.640	0.167	0.500	0.167	0.167	satisfactory
19	0.333	moderate	-0.094	0.000	0.333	0.333	0.333	eliminated or completely revised
20	0.917	Too easy	-0.308	0.000	0.083	0.000	0.917	eliminated or completely revised
21	0.667	moderate	0.377	0.250	0.000	0.667	0.083	little or no revision
22	0.417	moderate	0.204	0.000	0.417	0.167	0.417	need revision
23	0.583	moderate	0.388	0.083	0.333	0.583	0.000	little or no revision
24	0.917	Too easy	0.446	0.917	0.083	0.000	0.000	satisfactory
25	0.583	moderate	0.420	0.083	0.333	0.583	0.000	satisfactory
26	0.750	easy	0.173	0.750	0.000	0.167	0.083	eliminated or completely revised
27	0.917	Too easy	0.171	0.917	0.000	0.083	0.000	eliminated or completely revised
28	0.750	easy	0.682	0.083	0.167	0.750	0.000	satisfactory
29	0.833	easy	0.470	0.083	0.083	0.833	0.000	satisfactory
30	0.250	difficult	-0.285	0.167	0.333	0.250	0.250	eliminated or completely revised
31	0.667	moderate	0.755	0.083	0.667	0.000	0.250	satisfactory
32	0.750	easy	-0.068	0.000	0.167	0.750	0.083	eliminated or completely revised
33	0.917	Too easy	0.446	0.083	0.917	0.000	0.000	satisfactory
34	0.500	moderate	0.319	0.250	0.167	0.083	0.500	little or no revision
35	0.167	difficult	0.086	0.333	0.167	0.417	0.083	eliminated or completely revised
36	0.417	moderate	0.298	0.250	0.417	0.250	0.083	need revision
37	0.333	moderate	0.489	0.167	0.167	0.333	0.333	satisfactory
38	0.167	difficult	0.498	0.167	0.500	0.167	0.167	satisfactory
39	0.167	difficult	0.540	0.083	0.333	0.417	0.167	satisfactory
40	0.833	easy	-0.177	0.833	0.167	0.000	0.000	eliminated or completely revised