Evaluating Indonesian Local Culture-Based English Reading Materials Evaluation Instrument: Preliminary Delphi

Ikhsanudin

{ikhsanudin@fkip.untan.ac.id}

Universitas Tanjungpura, Jalan Profesor H. Hadari Nawawi, Pontianak – Indonesia

Abstract. Cultural contents in language teaching are indispensable. One of the practical ways to include local cultures in language teaching is by developing and evaluating them properly using valid and reliable instruments. However, until the present time, at least in Indonesia, not many thoroughly evaluated instruments are available to support the evaluation. This article reports an evaluation of a copyright Indonesian local culture-based English learning materials evaluation instrument called Ipmabibul using the preliminary Delphi (pre-Delphi) technique. Besides reporting the evaluation, this research also introduces the evaluation technique. The pre-Delphi technique is designed to help instrument developers to obtain experts‘ agreement on the need to include certain evaluation items without agreeing on the formulation of the items. This research concluded that most of the items in the Ipmabibul could be included in the instrument, and some of them needed to be revised. The research also indicated that the pre-Delphi was usable.

Keywords: Delphi, evaluation instrument; external evaluation; local culture; Pre-Delphi, reading materials

1 Introduction

Digital technologies and innovations in information and telecommunication have enabled interactions among cultures from different parts of the world. Explosions of digital reading materials and other resources influence the improvement of digital reading activities. A study in Africa indicated the positive effects of digital reading [1]. The vast amount of reading materials on the internet has given opportunities for young people to do reading extensively, which is very useful for improving their intelligence, particularly knowledge of grammar and vocabulary [2], and promoting cultural education [3].

Students and young people across the globe are getting more engaged in multicultural life. Culture will be one of the most crucial problems to solve, even in large countries. Recently, a study explained the importance of constructing a multicultural education, showed the possibilities for establishing multicultural education, and pointed out the paths to constructing

ICEL 2022, November 07-08, Malang, Indonesia
Copyright © 2023 EAI
DOI 10.4108/eai.7-11-2022.2329253

multicultural education in China [4]. One study in Finland advised that, in facing this global world, cultural traditions should be maintained through education [5]. A study in Oman reported that teachers and students were optimistic about promoting extensive reading programs even though they also knew that there were issues that impeded their optimism [6].

To make to support reading activities, it is vital that the targetted reading materials are available and accessible. In the Indonesian experience, the government invited experts to write books, procured reading materials, and evaluated the books before serving them to the students [7]. It is also possible that teachers can collaborate to write reading materials for their students as global or national materials, like what has been done by a student in Pontianak, an expert in Pontianak, and an expert in Pulau Pinang that collaborated in producing a set of teaching materials [8].

As an attempt to provide quality education, reading materials should be evaluated by experienced evaluators using quality evaluation instruments. To result in good evaluation, an evaluation instrument must be appropriately evaluated. An example of evaluating an instrument evaluation was done by Fiktorius, Ikhsanudin, and Salam. The researchers conducted a validation study on a set of national English examination test items in Indonesia [9]. Evaluation of instruments is very essential in the attempts to provide quality education and educational measures [10].

This research was conducted to evaluate a set of checklist items of the instrument of local culture-based English language materials evaluation. The instrument is listed in the Ministry of Law and Human Rights of the Republic of Indonesia, Number EC00202052439, dated 24 November 2020, as *Instrumen Penilaian Materi Bahasa Inggris Budaya Lokal (Ipmabibul)*, but it is not supported by any research document. This research is vital to provide scientific findings that can be referred to by potential users of the instrument. The purpose of evaluating the instrument was to know the experts‘ acceptance of every item in the Ipmabibul.

The results of the evaluation are helpful in the process of revising the instrument. In turn, local culture-based English language materials that are available and that will be produced can be evaluated. The implication of this research is that more quality local culture-based English language materials will be available to support culture education.

This research has two novelties. The first novelty is that this research evaluates an instrument that has not been previously evaluated through a study that involved a significant number of experts. The second one is that this research offers a new way, which is simple and valuable, of evaluating evaluation instruments. What is called new in the way of evaluating is that the researcher modified the Delphi method, which is only the preliminary stage before implementing the established Delphi. This method is called Preliminary Delphi or Pre-Delphi technique. This technique also can be called Quantitative Delphi because the data that are collected and analysed in this study are quantitative. The name Descriptive Delphi is also suitable because this Delphi tends to describe what is said by the experts, particularly those who are involved in the evaluation process.

The Delphi method is a way of obtaining consensus on a particular subject using circles of questioning that involves experts in the relevant research interest. In this version, Delphi is defined as having three characteristics: (1) the responses and interactions are within an anonymous group; (2) the questionings happen in multiple cycles; and (3) the provision of feedback to the group is between each cycle [11]. The biggest problem among the three characteristics is the multiple cycle questioning. Multiple cycles procedure usually takes time and energy.

In this research, the researcher proposes the preliminary Delphi as follows. The experts�flesh responses are analysed using descriptive statistics, namely, mean (μ) and standard deviation (std) or deviation (d). The mean score is the representation of a dataset‘s average value. In statistical calculation, the mean score is essential because it informs the researcher about the position of the centre of a dataset [12]. Whereas standard deviation is substantial because it informs the researcher about how spread out the given dataset values are [13]. The combination of the mean and the standard deviation provides information to the researcher on the trend of the observation score, which in this research means the trend of the experts‘ preference or acceptance can be found. To minimize bias in gaining collective responses—such as peer and societal pressures—the responses are kept anonymous while the number of experts involved in the study is enlarged.

A recent study reported that library and information science (LIS) Delphi were not frequently studied. However, most of them were published in the most prestigious academic journals. Out of 105 Delphi research articles between 1971 and 2019, there were at least ten variants of Delphi techniques, namely: Classical e-Delphi (30 articles), Modified e-Delphi (23 articles), Classical Delphi (17 articles), Policy e-Delphi (9 articles), Modified Delphi (9 articles), Critical e-Delphi (7 articles), Policy Delphi (4 articles), Critical Delphi (2 articles), Grounded Delphi (2 articles), and Online Delphi (2 articles) [14]. This new Dephi is proposed as a new alternative to implementing the Delphi method; to provide a solution to the complexity of the “unlimited rounds/cycles“ that may happen in the most Delphi method. One of the weaknesses of this new Delphi, however, is that this method cannot guarantee that the evaluation will be completed in one round.

2 Method

The method used in this study was evaluation research, particularly the external evaluation of an instrument. To do the evaluation, this study used the Delphi technique. The technique that was implemented was Pre-Delphi, which is a new variant of the Delphi technique that is first introduced in this article. Three principal components were involved in this evaluation research, namely: the researcher that also acted as the evaluator, the instrument to be evaluated, and the experts that gave responses to every item in the instrument.

The purpose of using Pre-Delphi is to obtain English language teaching experts‘ consensus on whether the newly designed questionnaire items in the Ipmabibul may be used to evaluate local culture-based English language materials. The researcher asked questions to the experts if each item in the Ipmabibul could be included in the instrument without focusing on the sentence formulation of the items. The sentence formulation should be evaluated in another evaluation that uses “standard“ Delphi. By using Pre-Delphi, it is expected that the Delphi evaluation that will be administered in the next stage will take fewer rounds than without Pre-Delphi.

The implementation of this Pre-Delphi is described as follows. The data were obtained through structured interviews that involved 37 participants out of 60 experts who were invited to participate. The experts were professors, associate professors, and assistant professors in the field of English language teaching who worked in nine universities in four of five main islands in Indonesia, namely Kalimantan, Java, Sumatra, and Sulawesi. The data of this research are the experts‘ responses to the researcher’s questionnaire about each item of Ipmabibul. The main question says: “Is each of these items essential to being included in the Ipmabibul?“ The respondents are given four choices of the Likert scale options: (1) really unimportant, (2) unimportant, (3) undecided, (4) important, and (5) really important. An expert’s choice of a particular questionnaire idem indicated his/her acceptance of the item but not an indication of accepting the format or the sentence formulation of the item.

The analysis that was used in this study was descriptive statistics to obtain the highest scores, the lowest scores, the means, and the standard deviation. Those data were used to determine the tendency of the lecturers’ perceptions about every item or evaluation question about the English reading materials. The analysis was also conducted to find the deviation standard of every questionnaire item. Using the mean score and deviation of each item, the researcher categorized the lecturers’ preferences for the item, as represented in Table 1.

Table 1. Model of Categorizing Questionnaire Items Based on Preferences

.	High Deviation (HD)	Low Deviation (LD)
High Mean (HM)	Recheck for revision (B)	Accept for use (A)
Mid Mean (MM)	Parish (D)	Consider the urgency (C)
Low Mean (LM)	Recheck for revision (B)	Perish (D)

Note: This model is first proposed in this article

Table 1 can be explained as follows. Category A means the item can be used without revision because the item obtains high respondent preference with low deviation. Category B means the item should be rechecked for revision because it obtains the majority of the respondents’ preference, but some respondents rated it low. Category C means the item should be reconsidered whether it is really urgent to use because most of the respondents seemed unsure. Category D indicates that the item should perish because most respondents seemed doubtful with high deviation, or all of the respondents rated it low with insignificant deviation. In the next step, the result of calculation and categorization were analyzed and evaluated.

The analysis and evaluation were done mainly on the items that were categorized as B and C. Meanwhile, the items categorized as A were directly accepted, and the items categorized as D were directly discarded. The analysis of B items was focused on grammar and wording. The analysis of C items was focused on the contents. The evaluation process was making a decision on whether an item was to be accepted or to be discarded. The B items were accepted when there was no issue with grammar and wording, and the C items were accepted when there was no issue with the contents.

3 Findings

The Ipmabibul consists of 37 items that are divided into six groups, namely language (6 items), cultural contents (11 items), printed instrument appearance (8 items), electronic instrument appearance (8 items), and graphic quality (4 items). The findings of this research are presented in tables. Every table below exposes the trend of the experts‘ acceptance of a group of items that may contribute to the quality of the instrument being evaluated. Every table exposes the minimum score (Min.), the maximum score (Max.), the mean score (μ), the standard deviation (d), and the evaluation category (EC) of the item in the evaluation that refers to Table 1.

3.1 Experts‘ acceptance of language and expressions

As exposed in Table 2, it is evident that experts responded that language and expression items in local English materials are essential to be included in the Ipmabibul. The items of grammar, diction accuracy, local terms maintenance, text logical acceptability, and text difficulty level obtained high mean scores and low deviation standards. In contrast, the native-likeness item should be rechecked for revision because it obtained a high standard deviation.

Table 2. Experts' acceptance of the language of the instruments (N=37)

	Min.	Max.	M	d	EC
1. The quality of the English written expression	2	5	4.66	.701	HM.LD=A
2. The English word choice accuracy	2	5	4.72	.634	HM.LD=A
3. The native-likeness of the English expressions	2	5	4.19	.896	HM.HD=B
4. Selective maintenance of local terms	3	5	4.66	.545	HM.LD=A
5. Logical acceptability of the English passages	3	5	4.56	.669	HM.LD=A
6. Difficulty level of the English passages	3	5	4.34	.787	HM.LD=A

Likert scale: 1 – really unimportant; 2 – unimportant; 3 – undecided; 4 – important; 5 – really important

It is interesting to highlight, however, the minimum and maximum scores in the first three items have relatively high differences. The high standard deviation scores in most items suggest that the data are more spread out than the following three items. That also means that, even though it is essential to include the quality of the grammar, the accuracy of the diction, and the native-likeness of the expressions in the evaluation of local English reading materials, view experts do not really agree with the inclusion of the first three items.

3.2 Experts‘ acceptance of cultural contents

The cultural contents of Ipmabibul consist of ten items. Most of them scored between 3 and 5.

Table 3. Experts' acceptance of instrument cultural contents (N=37)

	Min.	Max.	M	D	EC
1. Quantity or dominance of local-culture contents	3	5	4.44	.669	HM.LD=A
2. Passage's ethnic nuance in materials presentations	2	5	4.02	.803	HM.LD=A
3. Accuracy of local contents in the passages	3	5	4.59	.716	HM.LD=A
4. Accentuation of local culture's moral value	3	5	4.51	.621	HM.LD=A
5. Inclusion of local culture's way of life	3	5	4.27	.683	HM.LD=A
6. Inclusion of local culture's traditional households	3	5	4.24	.739	HM.HD=B
7. Inclusion of local ethnic technology and tools	3	5	4.24	.718	HM.HD=B
8. Inclusion of information on local cultural artefacts	2	5	4.10	.821	HM.LD=A
9. Inclusion of information about local cultural figures	2	5	4.39	.821	HM.LD=A
10. Inclusion of arts and games of local culture	3	5	4.32	.701	HM.HD=B
11. Inclusion of the history of local culture and community	3	5	4.29	.762	HM.HD=B

Likert scale: 1 – really unimportant; 2 – unimportant; 3 – undecided; 4 – important; 5 – really important

As exposed in Table 3, three items scored a minimum of 2, but all of the mean scores are above 4, which is considered high. However, some items obtained high standard deviations, which means the scores were not concentrated on 4 or 5. In these issues, some of the items were in the category of "recheck for revision" (B).

3.3 Experts' acceptance of instrument appearance

Ipmabibul is designed to be used for either printed or electronic English language reading materials. It provides eight questionnaire items for the evaluation of each instrument version with slight differences, particularly in relation to the nature of printed and electronic reading materials. The results of each evaluation are presented in Table 4 and Table 5.

Table 4. Experts' acceptance of the instrument's printed materials (N=37)

	Min.	Max.	(μ)	D	EC
1. Paper type appropriateness	1	5	3.98	.914	HM.LD=A
2. Paper size appropriateness	1	5	3.98	.740	HM.LD=A
3. Font appropriateness	3	5	4.20	.856	HM.HD=B
4. Font size appropriateness	3	5	4.24	.844	HM.HD=B
5. Test layouts	3	5	4.54	.564	HM.LD=A
6. Graphic/pictures layout	3	5	4.46	.622	HM.LD=A
7. Margin appropriateness	2	5	3.83	.782	HM.LD=A
8. Column size appropriateness	2	5	3.90	.694	HM.LD=A

Likert scale: 1 – really unimportant; 2 – unimportant; 3 – undecided; 4 – important; 5 – really important

Interestingly, the experts had different opinions about the need to involve items in each version. Of the eight items in the instrument for printed reading materials (see Table 4), six were accepted and only needed to be rechecked for correction. Whereas, in the instrument for electronic reading materials (see Table 5), there are only two were accepted, and the other should be rechecked for review. It is also interesting to notice that the mean scores in Table 4 tend to be lower than those in Table 5. It can be figured out that the deviation standards in Table 5 are much greater than those in Table 4. Two items in Table 4 have minimum scores of 1, but the evaluation category is accepted. It could happen because only one expert gave a score of 1 to each of the items.

Table 5. Experts' acceptance of the instruments' electronic presentation (N=37)

	Min.	Max.	μ	d	EC
1. Application type appropriateness	3	5	4.51	.665	HM.LD=A
2. File size appropriateness	3	5	4.51	.671	HM.HD=B
3. Font appropriateness	3	5	4.24	.792	HM.HD=B
4. Font size appropriateness	3	5	4.27	.762	HM.HD=B
5. Test layouts	3	5	4.39	.712	HM.HD=B
6. Graphic/pictures layout	3	5	4.44	.619	HM.HD=B
7. Margin appropriateness	2	5	4.00	.801	HM.LD=A
8. Column size appropriateness	3	5	4.05	.707	HM.HD=B

Likert scale: 1 – Really unimportant; 2 – Unimportant; 3 – Undecided; 4 – Important; 5 – Really important

3.4 Experts' acceptance of the graphic quality of the instruments

Both printed and electronic share the same items to evaluate. Table 6 shows the result of the evaluation of four items about the experts‘ acceptance of reading materials‘ graphic quality, namely the suitability of colour and passages, the appropriateness of pictures contents of texts, picture size suitability with text, and quality of pictures/graphics.

Table 6. Experts' acceptance of the graphic quality of the instruments (N=37)

	Min.	Max.	µ	d	EC
1. Suitability of colour and passages	3	5	4.71	.609	HM.LD=A
2. Appropriateness of pictures contents of texts	3	5	4.83	.491	HM.LD=A
3. Picture size suitability with text	3	5	4.32	.718	HM.HD=B
4. Quality of pictures/graphics	3	5	4.71	.554	HM.LD=A

Likert scale: 1 – really unimportant; 2 – unimportant; 3 – undecided; 4 – important; 5 – really important

Of the four items, item number three (picture size suitability with text) was categorized as "B", which means it should be rechecked for revision. Like other items, the mean score of the item is high, but the deviation standard is also high. The high deviation standards obtained by many tended to be the issue in this instrument.

4 Discussion

4.1 Ipmabibul's questionnaire items Quality

The evaluation findings say that, of 37 Ipmabibul questionnaire items, fourteen should be rechecked for revision. The first question that may be discussed about the findings is whether or not the fourteen items have a relationship that can be explained. It is systematic that the fourteen items obtained high mean scores and high deviation. The majority of the experts scored them high (4 and 5), but some experts scored them lower. As the theory says, the mean score indicates the position of the centre of a dataset [12]. Every item has a maximum score of 5 (highest). Two items obtained a minimum score of 1, eight obtained a minimum score of 2, and the others‘ minimum scores were 3. There was no item which obtained a minimum score of 4 or 5. It means the fourteen items that should be rechecked are potentially accepted. The problem may arise due to different interpretations by the experts/respondents on the sentence formulations of the items.

The second question is about data that were collected from the expert participants. Were the data low scores systematic? In other words, were the low scores given by the persons who have the same characteristics? To answer these questions, the researcher accessed the database or the raw data. It was found that the average scores given by different respondents vary from 3.5 to 4.9. Even though there was a tendency for certain experts to score higher than others, there was no tendency for certain experts to consistently give scores lower than others in every or majority of items. So, the height of the deviation, which tells how spread out the given dataset values are [13], is not caused by the respondents. The different interpretations of the instrument sentence formulations by the experts may not be caused solely by the experts‘ misinterpretations but also may be due to the quality of the sentence formulations. The fourteen items need to be revised to reduce the possibility of misinterpretations.

However, the existing 14 items cannot be rejected. The tendency of the data says that they can be accepted. The reason is that their mean scores are high, above the median and above the level of “undecided“ in the Likert scale range.

4.2 Pre-Delphi Usability

Can this research offer a piece of evidence that the pre-Delphi is usable? The pre-Delphi procedures that were implemented in this research have resulted in a set of data and analyses that are relatively obvious. The quantitative data of the experts‘ acceptance of Ipmabibul questionnaire items were organized and calculated in a spreadsheet and produced data analysis displays. The evaluation categories A, B, C, and D could help the researcher evaluate the instrument.

It is a fact that 14 out of 37 questionnaire items needed to be revised. Using mean scores and deviation standards, which are the main components of data analysis in the pre-Delphi, the researcher could detect that all the questionnaire items are appreciated by the evaluating experts. At the same time, the high standard deviations of the 14 items tell that the high mean scores are not convincing. The 14 items may contain expressions that cause misunderstanding or multi-interpretation and need to be rechecked for revision. The nature of quantitative data in the pre-Delphi leads the pre-Delphi research to objective analyses and findings.

4.3 Research Limitation

An instrument that is being developed needs intensive examinations and extensive tryouts as part of the process of evaluation. Either the Ipmabibul instrument or the pre-Delphi was first evaluated in this research. To reach the intended quality, they need more evaluations and revisions.

5 Conclusions and Recommendations

This pre-Delphi evaluation research concludes that, even though parts of the items need to be rechecked for revision, most of the items of Ipmabibul can be included in the next step of evaluation. The evaluation category of rechecking for revision does not suggest that the researcher exclude those items. The apparent results of the research steps show that pre-Delphi offers an alternative way that can help instrument evaluation be administered objectively before proceeding the evaluation to qualitative Delphi, which tends to be subjective.

It is recommended that more researchers contribute studies to develop more detailed and operational procedures of the pre-Delphi technique to help instrument developers and evaluators assess their research outcomes.

References

[1] Bouaamri, A., Otike, F., & Barátné Hajdu, Á.: Explosion of digital resources and its effects on the development of digital reading culture in Africa. Library Hi Tech News, Vol. ahead-of-print No. ahead-of-print. DOI: http://dx.doi.org/10.1108/LHTN-12-2021-0096

[2] Celik, B.: The role of extensive reading in fostering the development of grammar and vocabulary knowledge. International Journal of Social Sciences and Educational Studies. Vol. 6(1), pp. 215-223 (2019) DOI: http://dx.doi.org/10.23918/ijsses.v6i1p215

[3] Isnaini, M., Faizin, F., & Anisa, R. N.: Integrating cultural material as the media for extensive reading for teaching Bahasa Indonesia as foreign language. Satwika: Kajian Ilmu Budaya dan Perubahan Sosial. Vol. 5(1) 131-141 (2021) DOI: http://dx.doi.org/10.22219/satwika.v5i1.15851

[4] Jin, Y., Li, L., and Luo, S.: Chinese multi-cultural education: possibilities and paths. International Journal of Educational Management, Vol. 28(3) pp. 299-305 (2014) DOI: http://dx.doi.org/10.1108/IJEM-04-2013-0061

[5] Korhonen, R. (2021), "Diversity and Cultural Heritage in the Finnish Pre-Primary Curriculum", Gonçalves, S. and Majhanovich, S. (Ed.) Art in Diverse Social Settings, Emerald Publishing Limited, Bingley, pp. 247-262. DOI: http://dx.doi.org/10.1108/978-1-80043-896-520211016

[6] Al-Siyabi, M. S.: Promoting extensive reading culture in Omani schools. International Journal of Teaching & Education. Vol. 1(1), pp. 1-17 (2020) DOI: http://dx.doi.org/10.52950/TE.2015.3.3.003

[7] Ikhsanudin, I.: Providing electronic coursebooks for school learners nationwide: Indonesian experience. In D. S. Anshori, P. Purnawarman, W. Gunawan, & Y. Wirza (Eds.), Language, Education, and Policy for the Changing Society: Contemporary Theory and Research. pp. 124-140. UPI Press. (2020). Available at: https://www.researchgate.net/publication/349966535_Providing_electronic_coursebooks_for_school_learners_nationwide_Indonesian_experience (Accessed: 20 October 2022).

[8] Utami, I. W., Tazijan, F., & Rosnija, E.: Developing supplementary materials for descriptive text writing using Ispring Quizmaker. Journal of English LanguageTeaching Innovations and Materials(Jeltim), Vol. 4(1), pp. 88-101 (2022) DOI: http://dx.doi.org/10.26418/jeltim.v4i1.51770

[9] Fiktorius,T., Ikhsanudin, I., & Salam,U.: A validation study on national English examination of junior high school in Indonesia. Jurnal Pendidikan dan Pembelajaran Khatulistiwa (JPPK). Vol. 3(6), pp. 1-8 (2014). Available at: https://jurnal.untan.ac.id/index.php/jpdpb/article/view/5672 (Accessed: 21 October 2022).

[10] Garger, J., Jacques, P.H., Gastle, B.W. & Connolly, C.M.: Threats of common method variance in student assessment of instruction instruments. Higher Education Evaluation and Development, Vol. 13 No. 1, pp. 2-17 ((2019). DOI: http://dx.doi.org/10.1108/HEED-05-2018-0012

[11] Chalmers, J., & Armour, M.: The Delphi Technique. in P. Liamputtong (ed.), Handbook of Research Methods in Health Social Sciences. Springer Nature Singapore Pte Ltd. pp. 715-136 (2019) DOI: http://dx.doi.org/10.1007/978-981-10-5251-4_99

[12] Zach: Why is the Mean Important in Statistics? Statology. February 18th (2022). Available at: https://www.statology.org/importance-of-mean/ (Accessed: 20 October 2022).

[13] Zach: Why is Standard Deviation Important? (Explanation + Examples). Statology. August 21st (2022). Available at: https://www.statology.org/why-is-standard-deviation-important/ (Accessed: 20 October 2022).

[14] Lund, B.D.: Review of the Delphi method in library and information science research. Journal of Documentation. Vol. 76 (4), pp. 929-960 (2020). DOI: http://dx.doi.org/10.1108/JD-09-2019-0178