Examining the Validity and Reliability of Instruments for Measuring the Word-Recognition Accuracy: A Pilot Study

Kalsum Mohamed, Hamidah Yamat

[a] chombie6576@gmail.com
SK Sungai Bunyi, Pontian, Johor, Malaysia
Universiti Kebangsaan Malaysia, Bangi, Malaysia

Abstract

An instrument, either a self-developed, adapted or adopted, is the most important aspect of a research. In the main study, two instruments (1.Word Reading Test Kit and 2.A Reading Text) were adapted from other resources to assess the word-recognition accuracy. It is fundamental to test the validity and reliability of the instrument if it is self developed or adapted from any resources. The aim of this pilot study is to determine the validity and reliability of the two instruments for measuring the word-recognition accuracy. Two groups of samplings were selected. First, a group of 7 expert panels to get their feedback on face validity and content validity. The second was 8 Year 3 pupils to measure their scores for the test-retest reliability. The results showed that both instruments got 100% for 5 out of 8 criteria in the face validity which fell under the excellent category. For the content validity, both instruments obtained a high level of agreement among the 7 expert panels with (CVI =100%) for instrument 1 and (CVI = 98.57%) for instrument 2. In terms of test-retest reliability, the Pearson Correlation was r(8)=0.881 which indicated good reliability for the first instrument. Meanwhile, for the second instrument, the Pearson Correlation was r(8)=0.916. Thus, it indicated excellent reliability. After some minor amendment, the instruments were valid and reliable for data collection in the main study. A pilot study helped to identify the design flaws, gain experience and learn important information prior to undertaking the main study

Keywords: validity and reliability; instruments; measurement; pilot study

1. Introduction

Word-recognition accuracy must be acquired by young learners at an early age. Prior to a quantitative research of a quasi-experimental design of high frequency word games to enhance the word–recognition accuracy among Year 3 pupils in a rural school, two instruments were adapted from other resources to assess the word-recognition accuracy. Thus, a pilot study has to be carried out to test the validity and reliability of the instruments. A pilot study is carried out to see whether something can be done before a major study, and should the researchers proceed with the main study or not. It also asks how to carry it out. In simple words, a pilot study is paramount for the enhancement of the quality and efficiency of the main study. According to Arnold et al. (2009), a pilot study is the first step of an entire research protocol and is often a small-sized study, which helps in designing and modifying the main study. In addition, Ismail et al. (2018) state that testing a data collection instrument is important in ensuring the feasibility and consistency of the instrument in measuring an intended outcome of a research. This pilot study aims to determine the validity and reliability of the two instruments to be implemented for measuring the scores of the word-recognition

accuracy in the main study. For this pilot study, face validity and content validity are the key types of validity and test-retest is the type of reliability. It is conducted to sort out all the possible problems that might lead to failure of the main study. The procedures include the research time management during the COVID-19 pandemic.

2. Literature Review

2.1 Piloting Instruments for Validity and Reliability

Validity and reliability were very important aspects of quantitative research. According to Neuman (2003) there were multiple meanings of validity and reliability. He added that they were represented in many types or forms. In short, testing for the validity and reliability were crucial for researchers. Klenke et al. (2016) defined validity as the accuracy of an instrument in measuring the anticipated construct within a research. There were 7 key types of validity in a research: 1) face validity, 2) content validity, 3) Construct validity 4) Internal validity, 5) External validity, 6) Statistical conclusion validity and 7) Criterion-related validity. Face validity was the degree to see if the instruments appear to look like as what it was supposed to measure. Gelfand and Hartmann (1975 as cited in Azwani et al., 2017) put forward that for an inter-rater agreement, a minimal acceptable value of Kappa capitulated at 0.60 (60%) whereas Fleiss et al. (2003) stated that the percentage of inter-rater agreement yielded 70% [Kappa value = 0.70] fell under a fair to good category. On the other hand, content validity was the degree to see to what extent the items on the instruments were fairly representative of what it should measure. Content validity must be checked in the development process of an instrument to minimize any potential error associated with the instrument. With reference to Sangoseni et al. (2013), a CVI (content validity index) of 78% and above indicated a high level agreement if the rating was done by members of seven expert panels or more. On the other hand, Braun et al. (2019) defined reliability as the stability and consistency of scores from an instrument. There were 3 types of reliability: 1) internal consistency, 2) equivalent form and 3) test-retest. According to Price et al. (2015), a test-retest Pearson correlation of +.80 or greater, generally considered as good reliability. Test-retest reliability coefficients of stability varied between 0 to 1, where below +0.6 considered as questionable reliability.

3. Research Method

It would not be an easy task to carry out a big scale quantitative research during the COVID-19 pandemic among young learners. Thus, only one rural school in Johor state was selected by the researcher to conduct the major study and this pilot study. The instruments employed were used to assess the word-recognition accuracy among young learners. Both instruments focused on high frequency words that must be acquired accurately by young learners.

3.1 Samplings

In order to test the validity and the reliability of the instruments used in this pilot study, two groups of samplings were selected. First sampling was selected for the validity. It was a group of expert panels to get their feedback on face validity and content validity. The second sampling was selected for the reliability. It was a small group of Year 3 pupils to measure their scores for the test-retest reliability.

3.1.1 First Samplings

In order to review and rate the face validity and content validity of the two instruments, 7 expert panels were selected by the researcher. All of them were experienced academicians with minimum experience of 5 years and above in the field of teacher education and major in Teaching English as Second Language (TESL). From the 7 expert panels, two of them were ex LINUS2.0 facilitators, one was the SISC+ officer, one was the Head of English Panel of a rural school and the other three expert panels were English teachers.

3.1.2 Second Samplings

For the major study, the researchers had identified 23 samples from 31 pupils in a Year 3 class who met the criteria selection for a purposive sampling. Hence, the balances of 8 pupils from the same class were selected as respondents for the pilot study. The researchers had to choose only these 8 pupils as there were not enough pupils. In addition, it was not possible during this COVID-19 pandemic to get pupils from other classes who were willing to participate in this study. These 8 respondents would not be included in the major study to avoid bias. These 8 respondents were average and good readers.

3.2 Instruments to be piloted

The two instruments to be examined for validity and reliability were 1) Word Reading Test Kit and 2) A Reading Text. Word Reading Test Kit is an individual assessment of reading the 100 high frequency words accurately. For this Word Reading Test Kit, the researchers adopted and adapted the ideas from the Burt Word Reading Test (110 words) and from the Fountas & Pinnell Benchmark Assessment System (25 words). For the assessment of this Word Reading Test Kit, the respondents were allowed only 10 seconds to recognize each word and say it out. There were 10 words on a word list card. The researchers prepared 5 lists of 50 words for Form A and another 5 lists of 50 words for Form B. The 10 lists were student's copy while Form A and Form B were teacher's copy to record the score.

On the other hand, the reading text was a story from the collection of BBC Words and Pictures - Fun with Phonics entitled 'Tam's top hat'. It was a very short story with only 14 simple sentences which consisted of 12 high frequency words. The researchers chose this story because most of the words were repeatedly used in the story. In order to make it as a reading text instrument, the researcher had shortened the story to only 10 sentences, but still with those 12 high frequency words in it. Student's copy was in paragraph form while the teacher's copy was 10 sentences in table form. The researchers adapted some of the ideas in developing this instrument from the LINUS2.0 Screening for reading screening Construct 9 which was terminated by MOE in 2019. The rules for using this instrument were also adapted from it. For instance, each sentence carries one mark and only one error or mispronounced word in each sentence would be accepted. Respondents were allowed to read at their own pace.

3.3 Procedures of piloting the instruments

The researchers piloted the two instruments to examine the validity and reliability. In terms of validity, the instruments were given to the expert panels to get their review and rating for face validity and

content validity. In determining the face validity of the instruments, the responses of the expert panels were indicated by using "Yes" and "No" scale. The criteria of face validity of this study were as indicated below:

- The overall structure of the instruments in terms of construction and the format.
- The suitability of the format with young learners.
- The correct spelling of the words.
- The appropriateness of the font type and font size.
- The appropriateness of the amount of words/sentences.
- The appropriateness of the rules
- The appropriateness of the time allocated to read.

In addition, the expert panels were also requested to identify any deficient areas and provide either suggestions or recommendations on ways to polish up the two instruments. Meanwhile, in determining the content validity, the responses were indicated by using the "Agree" and "Disagree" scale. Agree denoted that the items in the instruments were relevant, laconic and needed minor amendment. In contrast, disagree denoted that the items in the instruments were either irrelevant or needed major amendment.

The meaning of reliability was the consistency of a measure. In terms of reliability, the two instruments were piloted for test-retest reliability as a measure of stability. It was conducted among the second sampling by measuring the scores of the 8 respondents for both instruments at two different times. The test-retest was carried out via WhatsApp or Telegram apps, as there was a Movement Control Order (MCO) due to the current COVID-19 pandemic. In between, no treatment or intervention was applied to the respondents. The researchers measured the scores for a second administration after 5 days.

3.4 Data Analysis

For data analysis, the responses of the 7 expert panels and the scores of 8 respondents were attributed to an excel worksheet and checked for any missing data values. The data were analysed in both qualitatively and quantitatively. For face validity the data was analysed based on the comments of the expert panels and the inter-rater agreement. Bowling (2009) stated that the coherence level of calculating an inter-rater agreement was using the percentage. Face validity was analysed using inter-rater agreement Cohen's Kappa Index (CKI) introduced by Cohen (2013). According to Wynd and Schaefer (2003), a perfect agreement between two or more raters was when the value of Kappa equal to +1 (100%). On the other hand, the content validity was analysed using Content Validity Index (CVI) as it was related to the degree of agreement among the expert panels.. The 'Agree' responses were assigned a score of +1 while the 'Disagree' responses were assigned a score of +0. As for the test-retest reliability, the data were computed to see the correlation between the two scores of each instrument. The data were analysed using the statistical measure Pearson Correlation Coefficient (PCC) which was also known as Pearson's r.

4. Results and Discussion

4.1 Face validity

The results and comments by the expert panels for 'Word Reading Test Kit' instrument were shown

in Table 1 and for 'A Reading Text' instrument were shown in Table 2.

Table 1: Result and Comments on Word Reading Test Kit for Face Validity

Criteria	Yes Responses	Percentage	Comments
Format	7	100	Acceptable. Good.
Suitability	7	100	Suits the primary school. 10 words per list is good.
Spelling	7	100	Good.
Font type	5	71.43	Use 'syazalina83v3' or 'WakNan'
Font size	3	42.86	Enlarge the font for student's copy
Amount of words	7	100	Good, but try not to test all at one time. / 50 words would be enough for a research./ Test only 25 to 30 words.
Rules	7	100	Acceptable. / Fair enough. / Don't tell students the time limit, it makes them nervous.
Time allocated to read	4	57.14	Do not set a time. / If student paused for more than 10 sec, then skip to next word. Give them a chance to read other words.

The result in Table 1 showed that five criteria obtained 100%, which fell under the excellent category. Most comments were good and acceptable except for the amount of words. Even though all the expert panels chose 'Yes' for this particular criteria, some of them suggested not to use all the 100 high frequency words to measure the score for a research. Some suggested to keep the same amount of words as the Fountas & Pinnell Benchmark Assessment System which was 25 words. The criteria 'Font type' was in a fair to good category. The font type used in the instrument was 'Century Gothic' that looked like 'syazalina83v3' and 'WakNan' font which were recommended in the comments column by both expert panels who chose 'Yes' and 'No' as the responses. The other two criteria were below the minimal acceptable value of Kappa which were the font size (Kappa value = 0.42) and time allocated to read (Kappa value = 0.57). Almost all expert panels recommended to enlarge the font size, especially for the student's copy (the 10 words in 10 lists). Three expert panels disagreed with the time allocated to read the words and most of them commented not to set a time limit for children to read the words in a list. All the comments were consolidated and analysed for amendment to the instrument for the major study.

Table 2: Results and Comments on A Reading Text for Face Validity

Criteria	Yes Responses	Percentage	Comments
Format	7	100	Good format. Acceptable
Suitability	5	71.43	Divide into two short paragraphs
Spelling	7	100	on to or onto?
Font type	5	71.43	Use 'syazalina83v3' or 'WakNan'
Font size	7	100	Good.
Amount of sentences	5	71.43	Add more sentences. / Try to shorten sentence no 6 and no 8
Rules	7	100	Fair enough. Good.
Time allocated to read	7	100	Just don't let them paused for long.

The result in Table 2 showed that five criteria obtained 100%, which fell under the excellent category and the other three criteria obtained 71.43% which were in a fair to good category. The expert panels' comments were consolidated and analysed for any amendment.

4.2 Content Validity

The CVI results of the expert panels for the items in 'Word Reading Test Kit' instrument were shown in Table 3 and for the items in 'A Reading Text' instrument was shown in Table 4.

Table 3: Results on CVI for the items in 'Word Reading Test Kit'

Item No	Item	No of Agreement	CVI
1	Word Reading Form A	7	100
2	List A/Form A	7	100
3	List B/Form A	7	100
4	List C/Form A	7	100
5	List D/Form A	7	100
6	List E/Form A	7	100
7	Word Reading Form B	7	100
8	List A/Form B	7	100
9	List B/Form B	7	100
10	List C/Form B	7	100
11	List D/Form B	7	100
12	List E/Form B	7	100
		Total = 1200/1200 (100%)	

The result in Table 3 indicated a high level of agreement (CVI =100%) among the 7 expert panels. Thus, the content validity of this instrument was valid.

Table 4: Results on CVI for the items in 'A Reading Text'

No	Item	No of Agreement	CVI
1	**The** sun **is up**.	7	100
2	**It is** hot.	7	100
3	Min sits **on the** hut.	7	100
4	Min **is** hot.	7	100
5	Tam hunts **in the** hut.	7	100
6	Tam spots **a** red top hat **and** drags **it to the** pond.	7	100
7	Tam hops **into the** hat **and** rests.	7	100
8	**The** hat hits **a** rock **and** sinks **in the** pond.	7	100
9	Tam hops **on to the** rock.	6	85.71
10	Tam **can get to** Min.	7	100
		Total = 985.71/1000 (98.57%)	

The result in Table 4 showed a high level of agreement (CVI = 98.57%) among the 7 expert panels. It indicated that the content validity of this instrument was valid.

4.3 Test-retest Reliability

The Pearson's r correlation SPSS output were shown in Figure 1 for 'Word Reading Test Kit' and Figure 2 for 'A Reading Text'.

Correlations

[DataSet0]

Correlations

		FirstScore	SecondScore
FirstScore	Pearson Correlation	1	.881**
	Sig. (2-tailed)		.004
	N	8	8
SecondScore	Pearson Correlation	.881**	1
	Sig. (2-tailed)	.004	
	N	8	8

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 1: Pearson's r for Word Reading Test Kit

The Pearson Correlation in Figure 1 showed r(8)=0.881. Thus, it indicated good reliability for the instrument 'Word Reading Test Kit'

Correlations

[DataSet0]

Correlations

		TestScore1	TestScore2
TestScore1	Pearson Correlation	1	.916**
	Sig. (2-tailed)		.001
	N	8	8
TestScore2	Pearson Correlation	.916**	1
	Sig. (2-tailed)	.001	
	N	8	8

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 2: Pearson's r for A Reading Text

The Pearson Correlation in Figure 2 showed r(8)=0.916. Thus, it indicated excellent reliability for the instrument 'A Reading Text'

4.4 Discussion

In accordance with the results of the face validity, content validity and test-retest reliability, some amendments were done to both instruments by the researcher for the main study. In conducting the pilot study to measure the scores of the two instruments, the researcher took almost seven sessions with the 8 respondents and each session took about 30 minutes. In conjunction, the researcher decided to abide by the comments of the expert panels on choosing and testing only 50 words for the 'Word Reading Test Kit' instrument for the main study and keep the original instrument for another research or innovation.

5. Conclusion

A few minor amendments were made to the 'Word Reading Test Kit' instrument such as in time allocated to read. The researchers removed and changed it as a rule where it would be counted as an error and skipped to read if the reader paused for more than 10 seconds. The researchers also enlarged the font size and

decided to use the font 'syazalina83v3' as it looked bigger and clearer. In addition, the researchers reduced from 100 words to only 50 high frequency words for the major study. An adjustment was made to the 'A Reading Text' instrument where only the font was changed to font 'syazalina83v3'. Overall, the value of this pilot study was clear when the researchers could identify some factors to be amended that could possibly have a negative impact on the main study. After the minor amendment, the instruments were valid and reliable for data collection in the main study of 'High Frequency Word Games to Enhance Word-Recognition Accuracy among Year 3 Pupils'. The pilot study did help the researchers to identify the design flaws, gain experience and learn important information prior to undertaking the main study. One of the strengths of both instruments was that both could be carried out either face-to-face or via online medium applications such as WhatsApp, Telegram, Zoom, Google Meet or Microsoft Team. Thus, if there were lockdowns due to the current pandemic, data collection could still be carried out.

Acknowledgements

The researchers would like to thank everyone who were a part of this pilot study and help in making this a success to be published.

References

Ardrey, J. 2008. *Tam's top hat BBC Words and Pictures* - Fun with Phonics: (Letters and Sounds Set 5). Pearson Education Limited.

Arnold, D. M., Burns, K. E., Adhikari, N. K., Kho, M. E., Meade, M. O., & Cook, D. J. 2009. The design and interpretation of pilot trials in clinical research in critical care. Critical care medicine, 37(1), S69-S74.

Azwani, M., Nor'ain, M.T., & Noor Shah, S. 2017. Evaluating the face and content validity of a Teaching and Learning Guiding Principles Instrument (TLGPI): A perspective study of Malaysian teacher educators. Geografia-Malaysian Journal of Society and Space, 12(3).

Bowling, A. 2009. Research methods in health. McGraw-Hill Education.

Braun, V., Clarke, V., Hayfield, N., & Terry, G. 2019. Thematic Analysis. In P. Liamputtong (Ed.), Handbook of Research Methods in Health Social Sciences (pp. 843–860).

Cohen, J. 2013. Statistical power analysis for the behavioral sciences. (Revised ed.). Academic Press.

Fleiss, J.L., Levin B., & Paik, M.C. 2003. Assessing significance in a fourfold table. Statistical Methods for Rates and Proportions, Third Edition, 50-63. John Wiley & Sons.

Fountas, I.C. & Pinnell, G.S.1996. Guided reading: Good first teaching for all children (p.424). Heinemann.

Gelfand, D.M., Hartmann, D.P., Cromer, C.C., Smith, C.L., & Page, B.C. 1975. The effects of instructional prompts and praise on children's donation rates. Child Development 46,980-983

Ismail, N., Kinchin, G., & Edwards, J. A. 2018. Pilot study, Does it really matter? Learning lessons from conducting a pilot study for a qualitative PhD thesis. International Journal of Social Science Research, 6(1), 1-17.

Klenke, K., Martin, S., & Wallace, J. R. 2016. Qualitative Research in the Study of Leadership. Emerald Group Publishing Limited. https://doi.org/doi:10.1108/9781785606502

Neuman, W. L. 2003. Social Research Methods: Qualitative and Quantitative Approaches 5th Edition: A and B.

Price, P.C., Jhangiani, R.S., & Chiang, I.A. 2015. Research methods in psychology- 2nd Canadian Edition. BC Campus.

Sangoseni, O., Hellman M, & Hill, C. 2013. Development and Validation of a Questionnaire to Assess the Effect of Online Learning on Behaviors, Attitudes, and Clinical Practices of Physical Therapists in the United States Regarding Evidence-based Clinical Practice. The Internet Journal of Allied Health Sciences and Practice 11(2), 1-13.

Thorpe, W.G. 1976. The Burt Word Reading Test. 1974 Revision. Manual. The Scottish Council for Research in Education.

Wynd, C.A, & Schaefer M.A. 2002. The osteoporosis risk assessment tool: Establishing content validity through a panel of experts. Applied Nursing Research 15(3), 184-188.