ISSN 2616-7433 Vol. 4, Issue 10: 80-85, DOI: 10.25236/FSST.2022.041013

Development and Validation of the Ability Test of Guessing Word Meanings from Contextual Clues (ATGWMCC)

Jiang Chao*

Department of Foreign Languages, Guangdong Vocational Institute of Public Administration, Guangzhou, China
**Corresponding author*

Abstract: *Although the strategy of guessing word meanings from context plays quite an important role in L2 vocabulary learning and reading comprehension, only a few tests have been developed to measure the ability of guessing word meanings from contextual clues. This paper aims to construct and validate an ability test of guessing word meanings from contextual clues (ATGWMCC) to fill the research gap. The participants of this ATGWMCC are non-English major graduates, from whose performances the items of the original test are revised and modified to be more reliable and valid. This paper will illustrate the development and validation of the ATGWMCC from the following aspects: the process of test construction; validation study of ATGWMCC; item revision and conclusion.*

Keywords: *Guessing Word Meanings, Contextual Clues, ATGWMCC, Non-English Major Graduates, Test Construction, Validation*

1. Introduction and Literature Review

According to Nation (2001) [1], the strategy of guessing from context has been considered one of key components of the vocabulary development. Nattinger (1988) [2] also holds that guessing from context is the most frequently used way of inferring the meanings of the unknown words. Several have been conducted on the relationship between the ability to guess word meanings from context and vocabulary learning as well as reading comprehension, but only a few tests have been developed to measure the ability of guessing word meanings from contextual clues (Schatz and Baldwin, 1986[3]; Haastrup 1991[4]; Stuart and Yosuke, 2013[5]). Given the significance of this ability and the scarcity the validated tests to measure it, the present test is developed and validated based on the research to fill this gap.

The Guessing from Context Test (GCT) developed by Stuart and Yosuke (2013) is made up of 60 questions based on 20 passages. For each target word, there are three questions: the first question is to choose the part of speech of the target word; the second question is to locate the exact word or phrase (context) that help guess the meaning of the target word; the third question is to figure out what the word means. Such format of items which consists of three questions each are considered in our development of the present test, but it is dismissed due to the lengthy item stems, which are undesirable in a good test.

Haastrup (1991) defines lexical inferencing in the following way: 'involves making informed as to the meaning of a word in light of all available linguistic clues in combinations with the learner's general knowledge of the word, her awareness of context and her relevant linguistic knowledge' (p.40). In the present study, only linguistic clues are included within sentence levels to avoid the lengthy item stems, as mentioned above. Based on Qiu's (2007) [6] book, the ability to guess word meanings from contextual clues to be measured including inference clues, experience clues, cause and effect clues, example clues and lexical relationship clues. The development of this test is based on this definition. present test is developed to test non-English major graduates' ability to guess word meanings from contextual clues, which can facilitate their major study whenever reading English literature is

2. The Process of Test Construction

2.1 The preliminary stage

As mentioned above, the target test takers are non-English major graduates, so the target words are mainly chosen from the tests of CET-4, CET-6 and TEM4. Based on Qiu's (2007) book, items are constructed and included in the present test by either modifying the stems or some options of the items. By referring to this book, 5 indicators are also extracted: the inference clues, the experience clues, the cause and effects clues, the example clues and the lexical relationship clues (the synonym and antonym are combined into one). Altogether 28 items are designed with nearly 5 items testing each indicator.

2.2 The pilot study

In order to make sure that test takers choose the keyed answers based on inferences instead of their previous knowledge of the target words and that the present test is of moderate difficulty, a pilot study is administered to 2 non-English major graduates, one is with relatively higher English proficiency who has successfully passed CET-6 and obtained a relatively high 75 points in the National Entrance Examination for graduates while the other graduate is with lower English proficiency just scraping through CET-4. After the pilot study, one item that is too difficult for both of them to guess is deleted. Another three items which are too easy for them to guess correctly are also deleted. Therefore, the final version of the present test consists of 24 items; the number of items measuring the inference clues, the cause and effects clues, the example clues, the lexical relationship clues, and the experience clues is 4, 3, 6, 5, and 6 respectively.

2.3 Test administration and data collection

The present test is delivered to non-English major graduates from nearly 30 different universities in China, and the majority of them are in the first year of their graduate study. A total of 80 test papers are retrieved. Participants are required to mark the target words if they have already known them before. If more than 50% of the target words are known by the participants, the test papers will not be included in later data analysis. So 7 test papers are excluded and the number of the test papers used for data analysis is 73.

3. Validation Study of ATGWMCC

According to Messick (1989) [7], test validity is conceived as a unitary notion including numerous aspects which contribute to acceptable test behavior. However, in the present paper, the section will present two aspects of validity as separate entities (Schmitt, Schmitt and Clapham, 2001[8]) based on item analysis and factor analysis.

3.1 Item analysis

The results of the ATGWMCC have been analyzed with TAP and both the score reliability and the item quality are examined in the following.

3.1.1 General information and score reliability

The mean of the test scores is 15.438, which means that the students have chosen the correct answers for 64.3% of the total 24 items on average. The standard deviation is 4.121, which indicates how wide the range of distribution deviates away from the mean score. There are four estimates of the test's reliability using two different ways of split-half methods, KR20 and KR21. Given that split-half methods don't yield the unique estimate and KR21 assume that all items are equal in difficulty, the reliability coefficient given by KR20 is chosen for analysis. So the reliability coefficient of the test is 0.728, which indicates that 72.8% of the observed score variance is attributable to true score variance. Standard error of measurement from KR20 is 2.148, which indicates the discrepancy between the examinee's true score and observed score. Generally, the test score is basically reliable with an acceptable reliability coefficient.

3.1.2 Item quality

An overall evaluation of item quality could be made from examining item difficulty and item discrimination indices. The frequency distribution of item difficulties and item discrimination indices would be provided in the following.

As Table 1 shows, the range between the highest item difficulty (0.85) and the lowest item difficulty (0.41) is 0.44, which means the item difficulty of this test is relatively balanced. In addition, 75 %(18/24) of the items' difficulties are between 0.5 and 0.8, 8% (2/24) of the items' difficulties are above 0.80 and 16.7% (4/24) of the items' difficulties are between 0.40 and 0.50. Furthermore, 33.33% (8/24) of the items have the item difficulties (p=0.4-0.6) around the ideal value (p=0.5). The mean item difficulty is 0.643, which is also higher than the ideal p-value. From the frequency distribution of item difficulties, we can see that there are no extremely difficult items and only a few relatively easy items in this test, which means that in general the difficulty of the items is reasonable and only a few items should be revised.

Table 1: Frequency Distribution of Item Difficulties

Intervals	Frequency	Frequency rate
[0.40,0.45)	3	0.125
[0.45,0.50)	1	0.042
[0.50,0.55)	3	0.125
[0.55,0.60)	1	0.042
[0.60,0.65)	3	0.125
[0.65,0.70)	3	0.125
[0.70,0.75)	4	0.167
[0.75,0.80)	4	0.167
[0.80,0.85)	1	0.042
[0.85,0.90)	1	0.042

According to Ebel's (1965) [9] rule of thumb, 45.8% (25%+8.3%+12.5%) of the items function quite satisfactorily, for D is higher than .40 (See Table 2); 16.7% of the items should be eliminated or completely revised, for D is lower than .19; 12.5% of the items are marginal and need revisions, for D is between 0.20 and 0.29. The mean item discrimination (0.383) shows that generally the items are designed well, but some revision is still necessary. In addition, according to Table 3, only 12.5% of the items' point biserial correlation is less than 0.2, and these items are of bad quality. Based on the analysis of the two item discrimination indices, the items in this test are of good quality on the whole and only a few of them need revision.

Table 2: Frequency Distribution of Item Discrimination Index

Intervals	Frequency	Frequency rate
[-0.01,0.00)	1	0.420
[0.00,0.10)	0	0.000
[0.10,0.20)	3	0.125
[0.20,0.30)	3	0.125
[0.30,0.40)	6	0.250
[0.40,0.50)	6	0.250
[0.50,0.60)	2	0.083
[0.60,0.70)	3	0.125

Table 3: Frequency Distribution of Point Biserial Correlation

Intervals	Frequency	Frequency rate
[0.00,0.08)	1	0.042
[0.08,0.16)	1	0.042
[0.16,0.24)	1	0.042
[0.24,0.32)	5	0.208
[0.32,0.40)	3	0.125
[0.40,0.48)	6	0.25
[0.48，0.56)	5	0.208
[0.56,0.64)	2	0.083

3.2 Factor Analysis

In designing the present test, the contextual clues consist of five major types based on Qiu's (2007) book and each type of clues is test based on several items. In the following, the application of factor analysis to a confirmatory construct validation study will be illustrated. To test our main hypothesis that in our present test the contextual clues are divided into five types, a one-factor model is fitted to the data. Fit statistics indicates a relatively satisfactory fit of the one-factor model (See Table 4). As the first line of Table 4 shows, this model chi-square is statistically non-significant, which support the one-factor model. Although values of the RMSEA and SRMR are both less than .10 indicating good model fit, the upper bound of the RMSEA exceeds .10 and the value of CFI is not larger than .95. So there is a need to more closely investigate sources of poor model fit.

Statistics from another five lines show that these five clues are indeed tested by the corresponding items respectively, with CFI>.95, RMSEA<.10 and SRMR<.10. But the loadings for each clue are not very high, which indicates the impreciseness of the present model.

Table 4: Goodness of fit statistics for the one factor model

	n	χ2	df	CFI	RMSEA	90% Confidence Interval	SRMR
Gue	73	0.130	5	0.938	0.098	(0.000 0.208)	0.066
Cau	73	0.000	3	1.000	0.000	(0.000 0.000)	0.000
Exa	73	0.373	9	0.953	0.033	(0.000 0.138)	0.069
Exp	73	0.759	9	1.000	0.000	(0.000 0.092)	0.051
Inf	73	0.866	2	1.000	0.000	(0.000 0.121)	0.018
Lex	73	0.541	5	1.000	0.000	(0.000 0.146)	0.050

Notes: Gue=guessing from contextual clues; Cau=cause and effect clues; Exa=example clues; Exp=experience clues; Inf=inference clues; Lex=lexical relationship clues

4. Item Revision

Table 5: Illustrative Item Analysis Results on Item 3, 5, 7, 13, 16, 17, 20 and 23

	Item Responses (%)				Diff.	Index	Point Biserial
Item	A	B	C	D	p	Disc.	Corr.
3	9	13	20	30	0.41	0.30	0.24
5	11	31	11	19	0.42	-0.03	0.04
7	14	38	12	9	0.52	0.33	0.30
13	6	55	7	5	0.75	0.30	0.26
16	19	13	32	9	0.44	0.17	0.09
17	2	14	1	56	0.77	0.13	0.26
20	11	5	47	10	0.64	0.25	0.31
21	2	7	62	2	0.85	0.32	0.39
23	3	47	10	13	0.64	0.19	0.22

4.1 Extremely difficult items

From the frequency distribution of item difficulties, we can see that there are no extremely difficult items and only a few easy items in this test. So such items are examined again and revised.

As Table 5 shows, item 21) is a relatively easy item (p=.85). From the option distribution, it can be seen that this item has three nonfunctional foils (options 1, 2, 4). The inclusion of the three so obviously incorrect answers makes it easy for the less-able participants to choose the right answer, which therefore reduces the discrimination index. The options are therefore revised as follows (Option C is the correct answer): A. fussy B. annoying C. boring D. painful

21). After ten years in one job, Mike decided that his paperwork was humdrum and that his telephone sales job was dull.

A. workable B. tired C. boring D. interesting

4.2 Negatively discriminating items

According to Table 5, item 5) is a negative discriminator according to both the discrimination index and the point biserial correlation. With 45.5% (10/22 with one participant do not choose any options) of the low proficiency group and 43.3% (12/28) of the high proficiency group choose the right answer B, this item is a negatively discriminating item, which means that the option “carefully” is too easy for the low proficiency group, so perhaps it is better to replace “carefully” with a little more difficult word like “punctiliously”. Another point we need to pay attention to is that 28.6% of the high proficiency group chose the option D “patiently” which seems a little attractive to them, thus perhaps another word which is more irrelevant to the semantic meaning of the sentence can be used, such as “cozily”.

5). For an hour, the cat meticulously cleaned the kitten’s paws, face and body.

A. tenderly B. carefully C. slowly D. patiently

4.3 Seldom chosen distractors

With no participants from the high proficiency group choosing the option A in item 13), options A and C in 17), and option A in 23) as well as so few participants in high and low proficiency groups choosing the other options except for the keyed answer in all of the four items, it can be concluded that the right answer is too obvious and after checking the item difficulty of these four items it was found that these items are quite easy with item difficulty above 0.64 (as shown in Table 5). So the other three options excluded the keyed answer in item 13) may be changed to the following options: A. mean C. chary D. save, which all have something to do with “省钱的” which may attract the participants to choose. In the same way, the options in 17) can be replaced with A. doubtful B. wondering C. puzzled. And the options in 20:A. thirsty B. pleased. The options in 23) can be revised as: A. low C. distracted D. fragmentary.

13). Carnegie was very frugal. Even though he earned little, he saved most of his money and lived on very little until he saved $ 10, 000 for the investment to make him rich.

A. scant B. economical C. extravagant D. luxurious

17). Children are usually inquisitive about things around them.

A. uninterested B. questioning C. afraid D. Curious

20). After a day’s hunting, Harold is ravenous. Yesterday, for example, he ate two bowls of soup, salad, a large chicken, and a piece of chocolate cake before he was finally satisfied.

A. exhausted B. rapturous C. starved D. greedy

23). The instructor failed the student because of his sporadic attendance record. Occasional attendance in class was unacceptable to the teacher.

A. acceptable B. infrequent C. frequent D. Failed

4.4 Too attractive distractors

In item 3), with 25% of the high proficiency group and 31.8% (even higher than the keyed answer which is only 27.3%) of the low proficiency group choosing option C, it is considered that option C as a too attractive distractor and a check of the item shows that the ambiguity in the options makes it difficult for examinees to choose. So option C should be revised to make it less attractive and can be changed as ‘striking’.

In item 7), there are the same numbers of participants who choose option B and C in the low proficiency group, so option C can be revised as “tasteless” or something to make it less attractive.

In item 16), there are more participants(40.9%) choosing option B rather than the keyed answer (36.4%) in the low proficiency group, which means option B distracts them too much, for the word “liability” has two meanings itself one is “responsibility”, another is “obstacle” while in this context, the latter meaning is taken. So perhaps it can be replaced with a word more irrelevant to the context such as “problems”. And there is another possibility that these students may refer to dictionaries to get the meaning for the reason that our testing is distributed to them in an electronic version, so their behavior can not be controlled.

3). The cryptic message made the agents wonder whether or not their code had been revealed.

A. hiding B. mysteries C. vague D. mysterious

7). The food at this restaurant is mediocre; you won’t rave about how delicious it is, but you aren’t likely to get sick from it.

A. tasty B. ordinary C. spoiled D. expensive

16). Two liabilities I see in moving to California are a higher cost of living and the fact that I will be so far from my family.

A. debts B. responsibilities C. obstacles D. dilemmas

From the above analysis of the item difficulty and item discrimination indices, it can be seen that this test is of good quality in general. Although Table 6 shows that the reliability coefficient would be improved if these four items (items 3, 5, 16,23) are removed, the problematic items are revised as shown in previous sections rather than simply removed(See Appendix B).

Table 6: The value of KR20 if the item deleted

Item	KR20 if Item Deleted
Item03	0.732+
Item05	0.747+
Item16	0.743+
Item23	0.732+

+ indicates that KR20 (0.728) improves if the item is removed

5. Conclusion

The present paper describes the development and validation of the ability test of guessing word meanings from contextual clues, targeting non-English major graduates. The ability of guessing word meanings from contextual clues is measured based on five subscales: inference clues, experience clues, example clues, lexical relationship clues as well as cause and effect clues. All together, a test of 24 items is developed. The results of 73 valid tests retrieved are analyzed with TAP and item difficulty and item discrimination indices are mainly analyzed to evaluate item quality. Generally, most of the items are of good quality and the reliability coefficient is relatively high. But there are still some problematic items, which are further confirmed by CFA. Although values of the RMSEA and SRMR are both less than .10 and the model chi-square is statistically non-significant all indicating good model fit, the upper bound of the RMSEA exceeds .10 and the value of CFI is not larger than .95. Further investigation of the sources of poor model fit is quite necessary. Based on item analysis and the results of CFA, some problematic items are revised, thus improving the reliability of this test.

Given the relatively small sample size, the small number of items, the deficient quality of some of the items and responses as well as our limited knowledge concerning construct definition and validation, the present test needs to be examined and improved by further studies. But it is not without merits. Basically, it can generally test the ability to guess word meanings from contextual clues. For those non-English major graduates who need to read a lot of English literature, it is of great value for them to measure their own ability in guessing word meanings and develop such ability, which would definitely improve their reading rate and facilitate their major study.

References

[1] Nation, P. (2001) Learning vocabulary in another language. Cambridge: Cambridge University Press.

[2] Nattinger, J. (1988) Some current trends in vocabulary teaching. In R. Carter & M. McCarthy (eds). Vocabulary and Language Teaching. London: Longman

[3] Schatz, E. K., & Baldwin, R.S. (1986) Contextual clues are unreliable predictors of word meaning. Reading Research Quarterly 21 (4): 439-53.

[4] Haastrup, K. (1991) Lexical Inferencing Procedures or Talking About Words. Tübingen: Gunter Narr.

[5] Stuart, A. W., & Yosuke, S. (2013) New Directions In Vocabulary Testing. RELC Journal 44 (3): 26277.

[6] Qiu, R. S. (2007) Skills for Guessing Unfamiliar English Words. Guangzhou: Sun Yat-sen University Press.

[7] Messick, S. A. (1989) Validity. In Linn, R. L., editor, Educational measurement. 3rd edn. New York: American Council on Education/ Macmillan Publishing Company, 1103.

[8] Schmitt, N., Schmitt, D., & Clapham, C. (2001) Development and exploring the behaviour of two new versions of the Vocabulary Levels Test. Language Testing 18 (1): 55-88.

[9] Ebel, R. L. (1965) Measuring educational achievement. Englewood Cliff, N. J.:Prentice-Hall.