TABLE ONE: THE RELIABILTY AND THE VALIDITY COEFFICIENTS FOR THE ORIGINAL AND THE FINAL CLOZE TESTS

Statistics	Original	Final cloze
Split-half, reliability, (odjusted)	.42	.91
KR-21	.35	.88
r_{xy}	.40	.85

the researcher thought it would be safer to compare the three sets of coefficients to determine of their difference were significat For this purpose, Hotelling's t-test was employed. The results of this comparison are reported in table 2.

TABLE TWO: THE t VALUES OF THE VALIDITY AND RELIABILITY COEFFICIENTS OF THE ORIGINAL AND THE FINAL CLOZE TESTS

Statistics	OC	FC	Difference	t Value
KR-21	.35	.88	..53	*10.16
Split-half	.42	.91	.45	*10.9
r_{xy}	.40	.85	.45	*7.77

* P<0.05

With the. 05 alpha level adopted and a two-tailed hypothesis employed, and *df* 79 the t value should exceed 1.98. This, therefore, shows that the difference in reliability and validity coefficients is significant due to factors other than chance with 95% certainty.

Conclusion

What has been found in this research project strongly propounds that the traditional item analysis techinque is a successful way of imporoving a cloze test characteristics. i.e. validity and reliability.

Now, if this technique improves the validity

and reliability of cloze tests, all those English teachers are use the cloze procedure in their classes can use it with more certainty as a testing technigue

Bibliography

Anderson, J. Charles: Native and Nonnative Speaker Performance on Cloze Tests. Language Learning. Vol. 30, No 16:59-76, 1980

Bachman, L.F: The Trait Structure of Cloze Test Scores. TESOL Quarterly. Vol. 16, No. 1:61-8, 1980.

: Fundamental Considerations in Language Testing. Oxford: University Press,1990.

Testing. Oxford: University Press, 1990.

Brown, J. : A Closer Look at Cloze: Validity and Reliability. In Oller, John W. ed. Issues in Language Testing Research. Massachusetts: Newbury House, 1983.

:Cloze Procedures: A Tool For Teaching Reading. TESOL Newsletter. Vol. 10, No. 5, 1986

Farhady, H. Birjandi, P. and Jafarpour, A: Language Skills Testing from Theory to Practice. Tehran: SAMT, 1995.

Jonz, Jon: Cloze Item Types and Second Language Comprehension. Language testing. Vol.8, No. 9: 1-22, 1991.

Instrumentation and procedures

The measures employed in this study included six versions of a cloze test constructed on the basis of a single passage and a Michigan test of English language proficiency. The passage used for constructing the cloze test was a 317- word text which was thought not to arouse any special interest among the subjects.

The cloze tests were administered at three different stages. At the first stage, a cloze test, called the original cloze, with every 7th word deleted was administered. At the second stage, four other 40-item cloze tests were developed, with the same deletion rate, but the deleted words were not the same. These five versions of the cloze tests provided a total of 200 different items which were analyzed to determine their item characteristics, i.e. item facility and item discrimination. After the analysis, 40 items with higher facility and discrimianation values were selected and these items were used to make the final cloze version. The final cloze test was administered during week 12. It needs to be mentioned that the four 40-item cloze tests at the second stage were administered during week 6 to groups of 20 students who had been selected randomly from the original group. The Michigan test which served as the criterion measure to determine the validity coefficients was also administered twice, First it was administered along with the original cloze during the first week, and the second time it was administered along with the final cloze during thetwelfth week.

Analysis

After the first stage of administering the tests, the descriptive statistics for the original cloze and the Michigan tests were calculated. Then the validity and reliability was estimated through KR-21 and split-half procedures and the Pearson product-moment correlation formula was used to determine the criterion-related validity index.

At the second stage, item facility and item discrimination indicies were calulated for the items of the original cloze test and four other cloze tests.

The statistical analyses at the third stage concerned the calculation of reliability and validity indicies for the final cloze test, Again, the reliability was estimated through KR-21 and split-half procedures. The validity coefficient was estimated through the Pearson product-moment correlation formula.

To explore the possibility that the final cloze test might produce higher validity and reliability coefficients from the original and final cloze tests were compared using Hotelling's t-test for differences in correlation coefficients.

Results

As it is indicated in table one, the reliability and validity coefficients of the final cloze tests are considerably higher. This information alone was enough to affirm the idea that a cloze test with items of reasonable facility and discrimination indicies will yield higher validity and reliability values.

However,

performance, blank length, and number of items (Granny, 1972; Anderson, 1979, Shohamy, 1984, Bachman, 1985, Zohervandi, 1988, Jons, 1991; etc). According to Brown (1988), "the effectiveness of any or a given combination of these variables seems to be related to two parameters, namely, how well they contribute to the tests' capacity to center the scores of a given population (as reflected by the mean score) and disperse the scores around the center (as reflected by the standard deviation)". One of the less investigated factors which seems to have a strong impact on accomplishing the aforementioned parameters and consequently improving the reliability and validity of the cloze test is the role of traditional item analysis techniques.

It is this factor which was under investigation in this reseach project. The purpose of this study was to find out whether the traditional item analysis techniques affect the reliability and validity of cloze tests. This would be done through calculating the item discrmination and item facility indicies of the items of a cloze test. These two idicies account for the individual differences among the testees. Therefore, it was hypothesized that if these individual differences are bounded up, there might be a kind of contribution to the total variance or the reliability and as a result to the validity of the cloze test.

Research Design

Subjects

The subjects involved in this study were a group of 80 undergraduate students studying English as a foreign language at Tehran University. They were mostly senior students. Except for a few students with Turkish background, the students who participated in this study spoke Persian as their mother thongue. The whole group comprised 46 female and 34 male students.

Abstract

Through out the past few decades, many research studies have been carried out to probe into the validity and reliability of the cloze test. Many scholars have manipulated such variables as scoring methods, passage difficulty, blank lengths, number of items and deletion patterns to find out the procedures most efficient for developing cloze tests with reasonable validity and reliability indicies. In line with these series of studies, the present study was conducted to see if using item facility and item discrimination values as criteria for item selection in constructing cloze tests have any effects on the improvement of the validity and reliability of cloze tests administered to Iranian students.

Key Words: Cloze Test, reliability, validity item discrimination, item facility

Introduction

The word "cloze" seems to have been an intentional corruption of the verb "close" in the English language. It may remind the reader of the process of "closure" in Gestalt psychology, refering to the natural psychological tendency in human beings to fill in gaps in patterns. The cloze procedure is also said to be related to the notion of redundancy in a passage. Carroll (1982) stated that the linguistic cues in the same sentence of a missing word help the examinee to supply it.

The cloze test can be simply defined as any passage of prose with an appropriate length and certain level of difficulty from which every nth word has been deleted. To construct a cloze test, the test developer has to go through four steps: first he selects an appropriate text. To choose such a text, the test developer should know the language capabilities of his tesstees. After determining the examinees' level of competency, the test writer may use a readability scale to select a suitable text. The second step concerns deleting words from the chosen passage. In general, there are three ways of word deletion: fixed ratio pattern, rational deletion, and random deletion. Every one of these methods may be used with different objectives in mind. The third step is to determine the length of a cloze test. The length of a cloze test is directly related to the number of deletions. The Research findings indicate that cloze tests with 25-30 blanks, with every 7th word deleted are the most efficient ones in terms of reliability and validity. As the last step, some researchers have recommended that the first and the last sentences of the cloze passage be left intact. Farhady, Birjandi and Jafarpour (1994) note that cloze tests with the first and the last sentences left intact have proven to be more valid and reliable. However, Klare, Sinaike, and Stolurow (1972), cited in Oller (1979) contend that this process is not necessary, though there is no harm in it, either.

Concerning all types of language tests in general, and cloze test in particular, the single most important factor is to see how well a particular test fits a given proup of testees. To examine the effectiveness (validity and reliability) of the cloze test, during the past few decades scholars have scrutinized such factors as deletion patterns, passage difficulty, scorning methods, native versus nonnative

The Role of Classical Item Analysis Techniques in Improving the Validity and Reliability of Cloze Tests

G.H. Rostamlu
Imam Hossein University

چکیده

این تحقیق به منظور تعیین تأثیر روش های سنتی تجزیه و تحلیل سؤالات به طور مشخص در میزان پایایی و اعتبار آزمون کلز انجام شده است ، به عبارت دیگر ، هدف از این تحقیق ، ارائه جواب به این سؤال بوده که : «آیا مجموعه سؤالات استاندارد (از نظر روش های سنتی) در یک آزمون کلز می تواند پایایی و اعتبار آن را تعیین کند؟»

برای یافتن جواب این سؤال ، در هفته اول ، یک آزمون کلز چهار جوابی با تعداد ۸۰ نفر از دانشجویان زبان انگلیسی برگزار شد . پس از آن ، سؤالات این آزمون مورد تجزیه و تحلیل قرار گرفت و شاخص های مربوط برای تک تک سؤالات تهیه شد . همچنین شاخص پایایی و اعتبار آزمون اولیه محاسبه شد . در هفته ششم ، چهار سؤالی دیگر بر مبنای متن اولیه طراحی شد و با گروه های بیست نفری از دانشجویان شرکت کننده در تحقیق به اجرا درآمد . سپس سؤالات این آزمون ها مورد تجزیه و تحلیل قرار گرفت و شاخص های مربوط به آن ها معلوم شدند . بعد از انجام این کار برای ساختن آزمون نهایی ، آن دسته از سؤالاتی که در آزمون اولیه و چهار آزمون دیگر ، شاخص های بهترین و نزدیک به استاندارد را داشتند ، انتخاب شدند و از ترکیب آن ها یک آزمون ۴۰ سؤالی درست شد . این آزمون در هفته دوازدهم با انجام گروه شرکت کننده در تحقیق اجرا و شاخص های اعتبار و پایایی آن نیز محاسبه شد .

برای یافتن جواب سؤال تحقیق ، با استفاده از فرمول های مربوطه ، شاخص های پایایی و اعتبار آزمون کلز اولیه و آزمون نهایی با هم مقایسه شدند ، نتیجة آن مقایسه نشان داد که شاخص های پایایی و اعتبار آزمون نهایی به مراتب بالاتر بودند . از آن جا که تقریباً تمام متغیرها در این دو آزمون یکسان بودند ، می توانیم نتیجه بگیریم که تغییر حاصل در میزان پایایی و اعتبار آزمون نهایی به دلیل وجود سؤالاتی بود که از نظر سیستم سنتی تجزیه و تحلیل سؤالات ، شاخص های بالایی داشتند .

بر اساس یافته های این بررسی می توان به معلمان عزیز توصیه کرد : برای حصول اطمینان از پایایی و اعتبار آزمون های کلز ، بهتر است ابتدا از تک تک سؤالات گنجانده شده مطمئن شویم و سپس سعی کنیم آن هایی را بگنجانیم که از نظر سطح آسانی و سختی و از نظر تفاوت گذاری (discrimination) بین افراد ، شاخص های قابل قبولی داشته باشند .