Education and Science

Vol 39 (2014) No 176 447-455

Psychometric Properties of Science Items Comprehension Test

Güçlü Şekercioğlu[1], Nihat Bayat [2], Sinem Bakır [3]

Abstract

This study aimed to develop Science Items Comprehension Test (SICT), which can be used for the purpose of measuring to what extent students can understand the written content involving 8th grade science items of SBS (national placement test), and determine its psychometric properties. In this survey study, 259 eighth grade students enrolled in elementary schools in Antalya constituted the participants. SICT was developed with thirty reading comprehension items with science items of SBS implemented in 2012 and 2013. Field experts' opinions were asked and necessary corrections were made based on experts' feedbacks. In order to determine psychometric properties of instrument, exploratory and confirmatory factor analyses were carried out with data collected from the research group. In accordance with factor analyses and results related to item discrimination, nine items were excluded from instrument, and the instrument took its final form with twenty one items. Results of exploratory and confirmatory factor analyses revealed that within the scope of single-factor structure, construct validity is high for target characteristics to be measured. In addition, reliability of scores acquired with SICT is high with regard to internal consistency.

Keywords

Reading comprehension
Inference
Science achievement
SBS
Test development

Article Info

Received: 06.20.2014
Accepted: 10.14.2014
Online Published: 12.16.2014

DOI: 10.15390/EB.2014.3692

Introduction

In recent years, the failure of the students in large scale related to science international exams (PISA, TIMSS, etc.) and science national exams (SBS, LYS, etc.) has reached remarkable level. Studies conducted to find out the causes of this failure have focused on physical impossibilities, methodological mistakes, students' attitudes and motivation. The failure in science exam can be explained with these factors to a limited extent. Besides, it is important to keep in mind that the aforementioned opinion was formed by means of the assessment of exam results. In this case, it can be thought that this failure is not only related to science knowledge but also reading comprehension skills used while reading and answering science items.

[1] Akdeniz University, Faculty of Education, Department of Educational Sciences, Turkey, guclu_s@yahoo.com
[2] Akdeniz University, Faculty of Education, Department of Elementary Education, Turkey, nihatbayat@gmail.com
[3] Akdeniz University, Institute of Educational Sciences, Department of Science Education, Turkey, sinem-bakır@hotmail.com

In national and international large scale exams, items are addressed in the form of written texts. These exams generally consist of multiple choice items. Students are expected to understand written textual items and to find out the correct answer. It can be thought that when students do not understand the item or misunderstand the item, they are not able to find out the correct answer. Therefore, the reading comprehension skills are a kind of equipment that might be used to overcome the first obstacle between the student and the item. Then, the science knowledge could step in.

Reading Comprehension and Inference: Reading comprehension is a kind of skill developed with a large amount of and various kinds of knowledge. Language use, grammar, perceptual and cognitive processes constitute the primary knowledge related to reading comprehension (Keçik & Uzun, 2004). In a specific reading process, the act of establishing meaning takes place by using these elements. Prior knowledge and the act of interpretation play an essential role in the establishment of meaning as a result of the interaction between the individual and the text (Akyol, 2003; Wallence, 1993).

Reading is the act of discovering the meaning the writer intends to communicate. When the reader finds out information via a text, he associates it with other knowledge, expectations and experiences he/she has (Grabe & Stoller, 2002). Most of the time, the meaning intended to be reached in the text is not explicit. To achieve implicit meaning is possible with active participation of reader through the points the text indicates. In this sense, readers are expected to observe how the text is constructed, and associate experiences and textual information (Fountas & Pinnell, 2001).

In the reading process, reader is the subject of the act of understanding. The reader is supposed to have competence, field information, language and discourse knowledge, critical thinking, and inferencing skills (Sweet & Snow, 2002). Each of these skills is a facilitator of analyzing text. Especially, if exam items are considered, the main purpose is to discover implicit information in text. In turn, this situation is related to inferencing skills.

Chikalanga (1992) describes inferencing as a cognitive process the reader uses to find out the implicit meaning in a written text on the basis of the context of the text and the prior knowledge the reader has. Inferencing is utilized to guess the meanings of unknown words, determine context, form a frame for interpretation, foresee circumstances and describe discordant cases (Trabasso, 1981; Nicholas & Trabasso, 1980). Function of inference is to gain missing information by associating pieces of information given in the text. With this aspect, inferencing is a scientific process that is put into practice and completes understanding.

Words and syntactic structures used by the writer in the text do not directly submit the very meaning, but act as mediators for the meaning. In this case, in addition to perception of information on the surface of text and analysis of linguistic structures, inferencing is needed. Reader's comprehension of the message depends on constructing meaning by fulfilling these points completely. According to Shimizu (2005), inferences are drawn to create cohesion and to extend content. Content of text is extended to reach missing information. With this kind of an inference, reader's prior knowledge begins to be articulated (McKoon & Ratcliff, 1998).

Apart from extension inferences, connection inferences are used as well. Connection inferences take place through intra-textual relations. Chikalanga (1992) expresses that these connections provide logical links between units in text. These connections are requisite for meaningful reading (Cain, 2006). Because in one aspect meaningful reading means reaching implicit information by configuring pieces of information.

Items given to students in exams are in the form of textual integrity. The primary aim is to assess students' levels of knowledge and skills in a subject; therefore, so some information in items is removed intentionally. In this respect, distracters and stems can be thought as text format that considerably requires inferencing. Student's success in inferencing, a component of student's reading comprehension skills, is based on associating information given in items and prior knowledge correctly. Hence, students have to see and perceive explicit and implicit options in the text.

In this regard, the aim of the present studies to determine the psychometric properties of Science Items Comprehension Test (SICT) involving 30 items converted from 40 items in Science Test of SBS 2012 and 2013 exams. Accordingly, the study addresses the following research questions:

1. How is the factor design of SICT?
2. Is single factor structure of SICT confirmed?
3. What are the item discrimination levels?
4. What are the item difficulty levels?
5. What is the coefficient of internal consistency?

Method

In this study, a survey method in quantitative manner was adopted. This study was conducted with 259 eighth grade students in Antalya. Three schools were selected for implementation. This selection was made on the basis of rating of these schools at "Transition from Primary to Secondary Education" (TPSE) (TEOG-national placement test). These schools were relatively chosen by their scores in the Transition from Primary to Secondary Education exam. Each school was categorized by the success level "high", "middle" and "low". .51% of participants were female (n=132) and %49 of participants were male (n=127).

Data Collection Tool

30 of the 40 items in SBS Science Test applied in 2012 and 2013 to 8th grade students were transformed into multiple choice items for the reading comprehension test. Expert opinion was received about the suitability of changing science test into reading comprehension test, whether items could be answered separate from science knowledge, Turkish grammar, and suitability for principles of measurement. In this direction, two faculty members of Elementary Science Education, two members of Turkish Language and one from Measurement and Evaluation were asked for expert opinion. According to feedbacks of experts, the instrument was ready after necessary corrections were made on item stems and item alternatives.

SICT has 30 items with four alternatives. A correct answer is scored with "1" point and a wrong answer and an item without an answer are scored with "0". After the analysis carried out to determine psychometric properties of instrument, 9 items were decided to be excluded from test and the instrument took its final form with 21 items.

In addition, for implementation of SICT, necessary permissions were received from Antalya Provincial Directorate of National Education.

Analysis of Data

In order to determine the psychometric properties of SICT, validity analyses were performed. To get an evidence of construct validity first exploratory factor analysis (EFA) was carried out. Then, to determine whether the single factor structure acquired as a result of exploratory factor analysis was verified, confirmatory factor analysis (CFA) was applied. EFA is a frequently used technique to reveal the sources of observed covariance and variance. This technique is quite useful for the first steps of scale development (Jöreskog & Sörbom, 1993). EFA aims to explore factor or factors based on relations between variables (Tabachnick & Fidell, 2001). Researchers try to get information about measuring instruments and nature of factors via EFA (Crocker & Algina, 1986). On the other hand, one of overwhelming advantages of CFA is making suggestions on variety of fit indices to evaluate rapport between theoretically described model and data. According to the literature, although there is not exact consensus on which fit indices should be used to evaluate models, using more than one fit index together is proposed (Byrne, 1994; Hair, Anderson, Black & Tatham, 1998; Netemeyer, Bearden & Sharma, 2003; Schermelleh-Engel, Moosbrugger & Müller, 2003; Sümer, 2000). In the present study, chi-square test (χ^2), chi-square and degree of freedom proportion (χ^2/df), the root mean square error of approximation (RMSEA), standardized root mean square residual (SRMS), goodness of fit index (GFI), non-normed fit index (NNFI) and comparative fit index (CFI) were evaluated. In most of the studies, both EFA and CFA are used; as a matter of fact, using CFA is desired after EFA (Jöreskog & Sörbom, 1993).

According to Comrey and Lee (1992), factor loading of .32 level is qualified as "poor" due to explained variance of 10% (as cited in Tabachnick & Fidell, 2001). Within this scope, it is accepted that desirable level of factor loading is at least .32.

With the purpose of evaluating item functioning with proofs related to structure reliability, for item discrimination, item analysis was carried out between the lower groups of 27% and the upper groups of 27% and criterion was accepted as $p<.01$. In the case of scoring items as 1 and 0, item scores are normally distributed, artificial discontinuous and dichotomous. Therefore, as item discrimination index for artificial dichotomous item scores and continuous test scores, point-biserial correlation coefficient or biserial correlation coefficients, which are specific kinds of Pearson product-moment correlation coefficient, can be used. However, if item difficulty indices are closer to edge (to 0 or 1), point-biserial correlation coefficient is affected more than biserial correlation coefficient from item difficulty indices. For this reason, if item difficulty indices are near .50, point-biserial correlation coefficient will give more information; on the other hand, if item difficulty indices approximate to edges, biserial correlation will be useful to obtain more information (Atılgan, 2006; Baykul, 2010). In this context, to produce additional proof for item discrimination and based on .51 average difficulty, point-biserial correlation coefficient was computed; moreover, criterion for discrimination is specified as $r_{jx}>.30$.

In addition to validity proofs, K-20 internal consistency coefficient was computed in order to determine reliability of test scores.

Results

Results of EFA, CFA, item discrimination levels and item difficulty level that were carried out to find answers to first four research questions are summarized in Table 1.

Table 1. Finding Related to the Psychometric Properties of SICT

Item No	EFA[1]	CFA			Lower 27%-Upper27% Between Groups I.D.	Point-biserial C.C.	Item Difficulty
	λ^2	Standardized Coefficient	T-Value	Error Variance	(r_{jx})	(r_{jx})	(p_j)
12	.925	.57	11.23	.68	.000	.54	.47
14	.919	.60	12.77	.64	.000	.60	.57
30	.910	.61	13.12	.63	.000	.61	.49
26	.905	.55	10.44	.70	.000	.52	.58
27	.900	.58	11.41	.67	.000	.55	.47
25	.890	.56	10.85	.69	.000	.55	.55
21	.881	.54	10.20	.71	.000	.52	.50
24	.880	.49	8.85	.76	.000	.46	.53
16	.858	.49	8.88	.76	.000	.49	.57
17	.855	.53	10.05	.71	.000	.56	.65
18	.847	.53	10.24	.72	.000	.54	.54
5	.829	.49	8.92	.76	.000	.53	.60
10	.820	.44	7.73	.81	.000	.49	.77
2	.791	.44	7.76	.81	.000	.46	.55
3	.788	.41	7.39	.83	.000	.44	.75
4	.743	.35	6.08	.88	.000	.38	.73
20	.719	.42	7.16	.82	.000	.40	.51
7	.704	.40	6.50	.84	.000	.43	.42
6	.661	.35	5.89	.88	.000	.41	.66
11	.484	.23	3.57	.95	.000	.25	.57
29	.347	.24	3.69	.94	.001	.22	.50

[1] Principle Component Analysis

In order to determine properties of factorial design, EFA was carried out. Before EFA, to test whether the sample size is sufficient for factoring, Kaiser-Meyer-Olkin (KMO) test was carried out. As a result of analysis, KMO value was calculated to be 0.831. In accordance with this finding, sample size can be acknowledged to be "sufficient" for factor analysis (Kalaycı, 2005). Moreover, results of Bartlett's Test of Sphericity revealed that chi-square value was found to be significant, $\chi^2(435)$= 1692.453, p=.000. Therefore, data was accepted as derived from multiple variable normally distribution. Also, measures of central tendency and coefficients of skewness and kurtosis were calculated, and data were found to be close to normal distribution. Moreover, there is not a multicolinarity problem between items.

As a result of EFA that was carried out depending on tetrachoric covariance matrix due to categorical scoring, it was identified that 30 items exploited in the analysis work under a single factor. It was determined that 9 items (1, 8, 9, 13, 15, 19, 22, 23, and 28) had a low factor load value (below .32). Before taking out these items, t values computed with CFA, error variances, item analysis between lower and upper groups of 27% and discrimination levels computed with point-biserial correlation analysis were evaluated together. Hence, the aforementioned items were left out of the scope of the test. As can be seen in Table 1, when these 9 items were excluded from the test, factor load values of the remaining items ranged between .347 and .925. Furthermore, it was found that single structure factor accounted for 64.98% of the total variance. The number of initial factors comprising 2/3 of the total variance related to variables included in the analysis is accepted to be the number of important factors. It is difficult to reach the aforementioned number in practice and especially scale development in behavioral sciences. For single-factor scales, a rate of 30% and above might be considered to be enough for explained variance (Büyüköztürk, 2014).

Results of the CFA performed to determine whether SICT's single factor design is verified reveal that t values for 21 items included in the analysis are significant. As illustrated in Table 1, standardized coefficients of indicators vary between .23 and .61, and error variances range between .63 and .95. When the modification suggestions were examined, it was decided to make a modification (between 25th and 26th items). It was found that contribution of modification to chi-square of modification is significant, p=.000. Due to categorical items, CFA was carried out based on asymptotic covariance matrices. Based on the fit indices derived from CFA, $\chi^2(188)=261.64$, p=.00031, $\chi^2/df=1.39$, RMSEA=.039, NNFI=.97, CFI=.97, SRMR=.054 and GFI=.90 values were computed.

Item discriminations of 21 test items were examined. According to the results of item discrimination analyses between lower and upper groups of 27%, it was found that all items are under the level of acceptance ($r_{jx}<.01$). Moreover, discrimination levels computed with point-biserial correlation technique vary between .38 and .61 except for the 11th and 29th item.

As seen in the Table 1, difficulty indices of items vary between .42 and .77; furthermore, average difficulty level of test was found to be .57.

Lastly, KR-20 internal consistency of SICT was computed as .86.

Discussion, Conclusion and Suggestions

Factor analysis is one of frequently used techniques to get evidence for construct validity in developing and adapting scales in social sciences and in research where a scale is used for a different aim or sample. Factor analysis is carried out to reveal factor structure or confirm estimated factor structure instead of offering a single validity coefficient of instrument. Data derived from factor analysis guide the validity and reliability analysis and other statistical analysis based on scores obtained from instrument. Factor analysis is a multivariate statistics used to discover limited number of new theoretically meaningful variables (factors/structures) by assembling a wide range of relevant variables or to test measurement models that explain relations between factors and indicators. There are two methods of factor analysis, namely explanatory and confirmatory factor analysis (Çokluk, Şekercioğlu & Büyüköztürk, 2012).

Proportion of explaining variance (approximately 65%) based on EFA and power of high factor loading values that explains relationship between items used to measure specific characteristics and items can be accepted as evidence of construct validity of SICT. Comrey and Lee (1992) state that in the case of .71 factor loading value, this value can be qualified as "excellent" due to explaining at least 50% of variance (as cited in Tabachnick & Fidell, 2001). In this context, factor loading values of 17 items of SICT could be accepted as "excellent" because these values are equal to .72 and higher than .72. 6th and 7th items of SICT can be identified as "very good" due to higher factor loading value than .63. 11th item of SICT can be identified as "moderate" due to higher factor loading value than .45. Moreover, 29th item of SICT can be identified as "poor" due to a higher factor loading value than .32.

In order to gain additional evidences concerning construct validity, CFA was carried out. Results of analysis reveal that fit indices meet the generally accepted level. In the literature, a χ^2/df ratio below 3 in large samples is regarded as a "perfect fit"; a GFI above .90 refers to "good fit"; a RMSEA below .05 is considered to be a "perfect fit"; a SRMR below .08 is evaluated as a "good fit"; and lastly a CFI and NNFI above .95 is regarded as a "perfect fit"(Brown, 2006; Byrne, 1994; Hu & Bentler, 1999; Jöreskog & Sörbom, 1993; Keloway, 1989; Kline, 2005; Schumacker & Lomax, 1996; Thompson, 2004). In this context, results of CFA confirm SICT has one way structure due to the fact that fit indexes generally meet acceptance level.

When the results related to 21 items of SICT are examined, both findings of item analysis between upper and lower groups of 27% and discrimination levels computed with point-biserial reveal that item discrimination levels meet acceptance level. The 11th and 29th items that did not meet the acceptance level computed by means of the point-biserial correlation technique were not left out of the test as they had factor loadings meeting the acceptance level in EFA and provided significant t values in terms of standardized coefficients demonstrating the explanatory power of the latent variable and indicator.

According to Nunnaly and Bernstein (1994), a validity coefficient of .70-.80 might be considered to be adequate for studies. In this context, internal consistency coefficient of scores obtained with SICT' as .80 indicates that it is acceptable in terms of reliability.

When the results are evaluated in general, it can be stated that validity of the scores obtained through SICT and the reliability in terms of internal consistency are high. In the literature, there are many studies investigating the causes of student failures by using tests developed within the scope of specific fields (science test, mathematics test etc.) in large scale exams as well as in-class assessments. Some of these research studies focus only on reading comprehension skills among the reasons for failure (Göktaş & Gürbüztürk, 2012; Imam, Abas-Mastura & Jamil, 2013; Kolıć-Vehovec, Bajšanskı & Zubkovıć, 2011; Ural & Ülper, 2013). However, achievement related to a specific field has been tested with instruments that aim to measure reading comprehension in native language. In order to determine the relationship between achievement in a specific field and reading comprehension properly, using reading comprehension tests consisting of the same item stems of the relevant field would yield more accurate results. In this respect, it is thought that SICT developed based on the stems of SBS Science Test will be an instrument that better serves the purpose of examining the relationship between reading comprehension and science achievement.

References

Akyol, H. (2003). *Türkçe ilkokuma yazma öğretimi*. Ankara: Gündüz Eğitim ve Yayıncılık.

Atılgan, H. (2006). Madde ve test istatistikleri. H. Atılgan (Ed.). *Eğitimde ölçme ve değerlendirme* (1. Baskı). s. 353-375. Ankara: Anı Yayıncılık.

Baykul, Y. (2010). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. (2. Baskı). Ankara: Pegem Akademi.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. (First Edition). New York: Guilford Publications, Inc.

Büyüköztürk, Ş. (2014). *Sosyal bilimler için veri analizi el kitabı: İstatistik, araştırma deseni, SPSS uygulamaları ve yorum.* (19. Baskı). Ankara: Pegem Akademi.

Byrne, M. B. (1994). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming.* California: Sage Publications, Inc.

Cain, K. (2006). Children's reading comprehension: The role of working memory in normal and impaired development. In Pickering, Susan J. (Ed.). *Working Memory and Education*, pp. 61-91, Amsterdam: Academic Press.

Chikalanga, I. (1992). A suggested taxonomy of inferences for the reading teacher. *Reading in a Foreign Language, 8*(2), 697-709.

Crocker L., & James A. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehartand Winston.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları*. (2. Baskı). Ankara: Pegem Akademi.

Fountas, I. C., & Pinnell, G. S. (2001). *Guiding readers and writers: Grades 3-6. Teaching comprehension, genre, and content literacy*. Portsmouth, NH: Heinemann.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading*. London: Pearson Education Longman.

Göktaş, Ö., & Gürbüztürk, O. (2012). Okuduğunu anlama becerisinin ilköğretim ikinci kademe matematik dersindeki akademik başarıya etkisi. *Uluslararası Eğitim Programları ve Öğretim Çalışmaları Dergisi (IJOCIS), 4*(2), 52-66.

Hair, J. F., Anderson, R. E., Black, W. C., & Tatham, R. L. (1998). *Multivariate data analysis: With readings.* Prentice-Hall International, Inc.

Hu, L., & Bentler, P. M. (1999). Cut off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Imam, O. A., Abas-Mastura, M., & Jamil, H. (2013). Correlation between reading comprehension skills and students' performance in mathematics. *International Journal of Evaluation and Research in Education (IJERE), 2*(1), 1-8.

Jöreskog, G. K., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Lincolnwood: Scientific Software International, Inc.

Kalaycı, Ş. (2005). Faktör analizi. *SPSS uygulamalı çok değişkenli istatistik teknikleri* içinde. Ş. Kalaycı (Ed.). (1. Baskı). Ankara: Asil Yayın Dağıtım.

Keçik, İ., & Uzun, L. (2004). T*ürkçe sözlü ve yazılı anlatım*. Eskişehir: Anadolu Üniversitesi Yayınları.

Kelloway, K. E. (1989). *Using LISREL for structural equation modeling: A researcher's guide*. London: Sage.

Kline, B. R. (2005). *Principles and practice of structural equation modeling.* New York: The Guilford Press.

Kolıć-Vehovec, S., Bajšanskı, I., & Zubkovıć, B. R. (2011). The role of reading strategies in scientific text comprehension and academic achievement of university students. *Review of Psychology, 18*(2), 81-90.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99*(3), 440-466.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications.* Columbia: Sage Publications.

Nicholas, D. W., & Trabasso, T. (1980). Toward a taxonomy of inferences for story comprehension. In F. Wilkening, J. Becker, & T. Trabasso (Eds.). *Information Integration by Children*. New Jersey: Lawrence Erlbaum Associates.

Nunnaly, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. (Third Edition). New York: McGraw-Hill, Inc.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit. *Methods of Psychological Research Online, 8*(2), 23-74.

Schumacker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling*. (First Edition). New Jersey: Lawrence Erlbaum Associates, Inc.

Shimizu, M. (2005). *Inference Generation Processes Of Japanese Efl Learners: Effects Of Questioning On Their Reading Comprehension*. Unpublished dissertation, University of Tsukuba: Japan.

Sümer, N. (2000). Yapısal eşitlik modelleri: Temel kavramlar ve örnek uygulamalar. *Türk Psikoloji Yazıları, 3*(6), 49-74.

Sweet, A. P., & Snow, C. (2002). Reconceptualizing reading comprehension. In C. C. Block, L. B. Gambrell, & M. Pressley (Eds.). *Improving Comprehension Instruction. Rethinking Research, Theory, and Classrom Practice*, pp. 17-54. USA: John Wiley & Sons, Inc.

Tabachnick, G. B., & Fidell, L. S. (2001). *Using multivariate statistics.* Needham Heights: Allyn & Bacon, Inc.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. (First Edition). Washington: American Psychological Association.

Trabasso, T. (1981). On the making of inferences during reading comprehension and their assesment. In J. T. Guthrie (Ed.). *Comprehension and Teaching*, International Reading Association.

Ural, A., & Ülper, H. (2913). İlköğretim matematik öğretmeni adaylarının matematiksel modelleme ile okuduğunu anlama becerileri arasındaki ilişkinin değerlendirilmesi. *Kuramsal Eğitim Bilim Dergisi, 6*(2), 214-241.

Wallence, C. (1993). *Reading (Language teaching: A scheme for teacher education)*. Oxford: Oxford University Press.