Psychology in Testing: Constructional and Conceptual Discrepancy in Language Tests

Sınavlardaki Psikolojik Unsurlar: Soru Yazma ve Soruyu Algılamadaki Uyuşmazlıklar

Mehmet ÖZCAN

Mehmet Akif Ersoy Üniversitesi, Eğitim Fakültesi, Yabancı Diller Bölümü, İngilizce Öğretmenliği Anabilim Dalı, Burdur/TÜRKİYE, mozcan@mehmetakif.edu.tr

ABSTRACT

This study investigates deficiencies in foreign language tests that stem from the discrepancy between test writers' assumptions and the test takers' approach to test items with a qualitative perspective. English language tests administered to elementary school students were analyzed. The analysis revealed that there are five types of common deficiencies. The deficiencies stem from the test items' requiring cognitive skills, general knowledge and higher developmental potential instead of linguistic abilities. Visual aids provided in the test items are not always useful for test takers and some test items require explicit grammatical knowledge which is not suggested by current teaching models.

Keywords: Foreign language testing, Constructional and conceptual discrepancy, Test item deficiencies, Validity, Reliability.

ÖZ

Bu çalışma, yabancı dil sınavı hazırlayanların, soruları yazarken sahip oldukları varsayımlar ile adayların bu sorulara yaklaşımlarında ortaya çıkan uyuşmazlıkları, nitel bir bakış açısıyla araştırmıştır. Ortaöğretim seviyesinde uygulanmış yabancı dil sınav soruları incelenmiş, bu sorularda, genel olarak, beş tür hata tespit edilmiştir. Bu hataların, soruların öğrencilerden, dil becerisi yerine, bilişsel beceriler, genel bilgi ve yüksek düzeyde gelişimsel birikim istemelerinden kaynaklandığı tespit edilmiş, bazı sorulardaki görsel ögelerin adaya katkısının olmadığı, bazı soruların ise güncel yabancı dil öğretim kuram ve yöntemleriyle çeliştiği gözlenmiştir.

Anahtar Kelimeler: Yabancı dil sınavı, Soru yazma-algılama uyuşmazlığı, Sınav hataları, Geçerlik, Güvenirlik.

INTRODUCTION

To paraphrase Bakhtin's (1981:280) understanding of dialogy, it can be proposed that any type of conscious human action is directed toward an ideal response and cannot escape the profound influence of the responding action that it anticipates. If this proposition is to be applied to test writing, every test writer has ideal test takers in her mind who are expected to respond to the test items written in the way the writer expects them to. However, because of human factor, there is almost always a discrepancy between the *ideal* and the *actual*; the discrepancy that is one of the greatest causes of discontentedness in human life. The discrepancy between the assumptions a test writer constructs in the process of writing each test item and the attitude, conceptualization and strategies the test taker develops in the approach to that test item is the cause of deficiencies in performance, validity, reliability and fairness of a test or a test item (see Fulcher and Davidson, 2007:62; Bachman, 1990:70). That is, the test writer himself turns out to be a factor in the evaluation of the test taker (see Withers, 2005). Since testing is a part of teaching and learning, this topic received attention from many researchers in the field of applied linguistics. Heaton (1990) outlines the relationship between testing and teaching and attempts to provide answers to *whats, hows,* and *whoms* of testing. He also explains general approaches to language testing in a comparative style (see also Bachman, 1990; Bachman and Palmer, 1996; Brown, 1996; Hughes, 2003; Izard, 2005) Differing from the previous literature, Weir (2005) presents a diachronic development in testing and analyzes some specific tests to demonstrate their structural features while discussing their power to test language as well. On theoretical plane, Xi (2010) attempts to bring a new conceptualization for the criteria that are used to evaluate the fairness of a test and proposes that fairness should be conceived as an aspect of validity (see Kane, 2010 for the criticism of this idea). Benedetti (2006) reviews the fundamental concepts in testing such as validity, reliability and fairness and demonstrates the stages of writing valid, reliable and fair tests. Baghaei (2011) investigated how C-test items should be prepared to be reliable and found that as the number of gaps in each passage increases item discrimination,

reliability and factorial validity of the test increase accordingly. Currie and Chiramanee (2010) investigated the effect of the multiple-choice item format on the measurement of knowledge structure. They found that there is a very strong relationship between test item format and what the test item measures. They claimed that multiple-choice items in a specific format fail to reflect the linguistic proficiency of the test taker and they maintained that if the test item format were to be changed, this change would urge the test takers mark other options than the one they have marked. Cunning-Wilson (2001) studied the use of pictures and other visual elements in testing and she suggests that the pictures or other visual aids should be appropriate to foster the understanding of test item psychologically and cognitively. She maintains that a visual element must also be digestible culturally. As an attempt to identify the sources of errors in language tests, Fazeli (2010) introduces a new type of reliability which he addresses as psychological reliability. According to his hypothesis, a test writer must take the test takers' decision making strategies into account while writing test items and these items should be piloted on a sample group of test takers to confirm that they are psychologically reliable. Perrone (2006) and Park (2010) investigated Differential Item Functioning, as it is defined by Subkoviak, Mack, Ironson and Craig, (1984) as 'Learners who have similar knowledge of the material on a test (based on total examination results) should perform similarly on individual examination items, regardless of gender, culture, ethnicity, or race'. Ironson, Homan, Willis, and Signer (1984) tested this hypothesis by 'planting' some unfair items into the test battery and they report that these items differentiate testers not on the skill and knowledge that is being tested but on some other criteria which the test was not intended to test.

The aim of the present study is to analyze certain test items with some degree of deficiencies in language test batteries that are administered throughout Turkey as achievement tests and in classrooms as progressive tests in order to identify the discrepancy between the assumptions of test writers while writing a test item and those of the test takers in the conceptualization of the same test item with a qualitative perspective.

METHODOLOGY

The sample test items were collected from language tests administered in past years to elementary school students either nationally or locally. The test items were analyzed and the ones that are identified as having defects were administered to 90 seventh graders in order to understand whether they will experience problems in answering them. Three American English teachers participated in our study as the normative testee group. We analyzed the construction and deconstruction processes of the test items with a more qualitative approach rather than quantitative one. Speaking, writing and listening tests were not analyzed. We assumed that all of the test items we included underwent a reliability and validity check by experts.

FINDINGS

The analysis of the test items reveals that there are mostly five types of problems that stem from the discrepancy between a test writer's assumptions in writing a test item and those of the test takers which are likely to decrease the achievement of a test taker in a language test.

Testing Linguistic Skills or Cognitive Skills

One of the fundamental concepts in testing is the term *validity* which is defined as the quality of a test's measuring only what it is intended to measure, but nothing else (Bachman, 1990:237; Brown, 1996:231; Heaton, 1990:159; Hughes, 2003:25). However, there were many test items requiring test takers to classify some words. Classification and categorization studies reveal that there are various conceptual and perceptual factors that influence the strategies children take in the categorization of an object or concept (see Gershkoff-Stowe, Thal, Smith and Namy, 1997; Rosch, Mervis, Gray, Johnson and Boyes-Bream, 1976). Since such tests tap the test takers' cognitive skills as much as their linguistic skills, their score is weak in reflecting the linguistic skills in a test taker. Test item (1) is extracted from a practice test.

(1) Aşağıda verilen sözcükler sınıflandığında hangisi dışta kalır?

(Which one of the following words is the odd one out if they are classified?)

A) mice   B) children   C) men   D) foxes

This test item is deficient for two reasons. First, test takers might approach each word in this test item according to their idiosyncratic experience or classification strategies relying on the nature of the object represented by the words provided in the options. For instance the size of the object, the frequency of the encounter with the object in daily life, the way the object is represented in fictional world would be factors affecting the choice of the test taker. Second, the test taker may approach the test item according to perceptual features the orthographic forms of the words present. Whether words in the options A, B and C are singular or plural does not influence the choice to be made by the test taker; simply, they do not contain the plural marker –s. The only one that contains it is option D. So, the right option must be D. If test takers come up with such a hypothesis, then the test item does not test the linguistic potential in the test taker. The choice would be made solely according to the existence or non-existence of explicit plural marker. In order to test our hypothesis, we administered this test item to 30 seventh graders. The findings show that 2 of the students (6.6%) marked *mice*, 6 students (20%) opted for *children,* 4 students (13.3%) marked *men* and 18 of them (60%) marked option *foxes* as the odd one out.

We restructured test item (1) and administered it to another group of 30 seventh graders in the same school. The words in test item (1) were replaced by the ones that are meaningless. The number of the letters and orthographic similarity were preserved in the nonsense words. The nonsense words are, *tice, klaytren, ven* and *cuxes* respectively. The instruction was exactly the same as the one given in the test item (1). It was observed that 2 students (6.6%) marked *tice,* 2 students (6.6%) marked *klaytren,* none marked *ven* and 26 of them (86.6%) marked the option *cuxes*. The analysis of the results of the two items implies two things: First, in both test items, the perceptual features of the words (e.g. –es in cuxes) had strong influence on the test takers' choice. Second, the proportion of the students who opted for the word containing the plural marker is lower

in the group who took the item containing meaningful words (60%) than that of the students in the group who took the item that contains nonsense words (86.6%). This implies that test takers in the former group classified the words not only according to the words' perceptual features but also according to the features of the referents the words in the test item evoke in the test takers mind. The semantic content of each word must have played a role in the choice of the test taker. As for the results obtained from the item that contains nonsense words, because the test takers were not distracted by semantic content of the words, they focused more on the perceptual features of the words in the test item. Thus, the proportion of the students who marked the option *cuxes* is higher than that of those who marked the option foxes.

Problems related to test writing do not appear only in English tests. The tests designed to test Turkish also present cognitive or psycholinguistic problems that are similar to those in test item (1). The instruction for test item (2) can be translated into English as "The antonym of which word among those underlined in the sentences below is **not hidden** in this word-search puzzle?"

(2) The antonym of which underlined words in the sentences A to D is not hidden in this puzzle?

A) **Alt** katın ışıkları yanıyordu.

   (The lights of the lower floor were on)

B) **Hafif** paketi ben taşıyabilirim.

   (I can carry the light package)

C) Sonunda **yoksul** günleri bitmişti.

   (Finally, his days of powerty were at an end.)

D) **Islak** havluyu balkona asıver.

   (Please, hang the wet towel in the balcony)

G	Ü	L	M	E	D	İ
A	S	İ	T	O	K	Ş
Ğ	T	M	A	S	U	M
I	D	O	S	T	R	A
R	E	N	İ	Ç	U	R

Table 1. The puzzle provided in the test item 2.

The antonyms for underlined words translate into English as follows:

    ALT: lower   X   ÜST: upper

    HAFİF: light   X   AĞIR: heavy

    YOKSUL: poor X   ***ZENGİN***: rich

    ISLAK: wet   X   KURU: dry

Solving a word-search puzzle does not require only linguistic skills (see Hambrick, Salthouse and Meinz, 1999; Larner, 2009). If the aim of the test containing item (2) is to test cognitive strategies along with the linguistic skills, this item may be valid (see Bachman, 1990:237). If this test item is intended to test only linguistic knowledge but nothing else, then it is unfair. Another flaw in this item is the target of the item: The testees are asked to identify not the word that exists but the one that does not exist in the puzzle. If the target were the former, the testees would match the underlined words with their antonyms in the puzzle and the task would end. If this were the task for the testee, the language tester would still have his will. Nevertheless, in the actual task in item (2), some of the testees are likely to scan not only the columns and rows but also the diagonal rows or reverse directions, in scrutiny, to be sure that the antonym of the intended word is not hidden there while others would stop the search right after they identify three of them. Another problem with the search in the puzzle is related to synonyms for the word *yoksul* (poor). When we look up the antonyms for the word *yoksul* in the online antonyms dictionary prepared by TDK (Turkish Language Institution), two words appear: *varsıl* and *zengin*. The task of searching in the puzzle for the test takers who possess both words in their mental lexicon will double.

Testing Topical Knowledge or Language

Test item (3) was extracted from a practice test that is administered throughout Turkey..

    (3) Murat- If you think analytically, then you are a left-brained person.
    Ali – Are you a left-brained or right brained person?

Murat- I am a right brained person because I think ....

A) rationally   B) analytical   C) sensitive   D) intuitively

We analyzed the test booklets of 43 students. 15 students (34.8%) chose option A: *rationally,* 1 student (2.3%) chose B: *analytical*, 8 of them (18.6%) chose the option C: *sensitive,* 19 students (44.1 %) chose D:*intuitively*. The booklets reveal that 31 of those who chose either A or D (72%), crossed out the options B and C. 23 of those who crossed out options B and C (74.1%) put a mark next to the option they finally eliminated and circled or ticked the other one as their choice, which implies that they considered both A and D as possible ones before they made their final decision. The fact that 72% of the students crossed out options B and C implies that these students are aware that the slot in the test item requires an adverb because the phrase to be modified here is a verb phrase: *I think*.

Another fact that 74.1% of the students who crossed options B and C spent some time on either option A or D to make their final decision hints that their decision is based on either the semantic scope of the words *rationally* and *intuitively,* or the truth value of the proposition about being a left or right brain person. If their decision is based on the former, their choice is still based on their linguistic skills. However, if the latter is the cause of hesitation for some of the students in choosing either A or D, then this test item is invalid. We administered the same test item to three American English teachers who teach English in an ELT department. Interestingly, 1 teacher chose option A while other two chose option D. This proves that test items requiring general knowledge test other skills more than language.

The Age and Content Discrepancy

Conceptualization of entities in external world is refined with increasing age because of both cognitive maturation and the semantic network constructed through personal experiences of children (see Piaget, 1929: 194; Piaget, 1930:237). Younger children tend either to overgeneralize or undergeneralize what they experience in their own environment (see Gershkof-Stowe, Connell and Smith, 2006; Pinker, 2004). Hence, the

relationship between test items and the age of the test taker is very strong. If a test item requires knowledge in some topics that is beyond the test taker's conceptualization, such test items would be invalid. Test item (4) is taken from a progress test that is administered to fourth graders.

(4) Aşağıdaki ülke isimlerini yanda verilen kelimelerle eşleştiriniz.

1. Netherlands   a. rice

2. Greece   b. hamburger

3. Italy   c. yoghurt

4. The USA   d. macaroni

5. Turkey   e. cheese

Although this test item requires the test takers to possess the semantic content of the words provided, the requirement for the knowledge that would help the test taker about which country is related with which food is stronger than knowing these words in English. One of the fourth graders dropped the following note at the bottom of the test item "Turkiye için 'ceviz ezmesi' olması lazım ama o da şıklarda yok (There must be the option "smashed walnut" for Turkey but it is not available among the options)" This shows that the test taker conceives of *yogurt* to be so ordinary to make a country "famous" and there must be something extraordinary such as "smashed walnut", for which the city in which the test taker lives is famous. It is the overgeneralization that prevents the test taker from choosing the "right option" in the mind of the test writer.

The Function of Visual Aids In Test Items

Providing background schemata can foster both assimilation process of input in learning situations and responding to a test item in a language exam (Canning-Wilson, 2001). Visual components in a test item must be inseparable part of the test item itself in that when the visual components are removed, the test taker must not be able to communicate with the test item or at least the communication must be too weak to ease the cognitive processing related to the item. Test item (5) is an item administered in a

city-wide practice exam. The fostering influence of the picture in test item (5) is open to discussion.

(5) Soruda verilen resme göre boş bırakılan yere gelecek uygun seçeneği bulunuz.

Maria's birthday is ....

A) in the thirteenth of June.

B) on the thirteenth of June.

C) on the thirtyth of June.

D) in the thirtyth of June.

The picture provided has nothing to do with the processing of the linguistic part of the item. Thus, let us put its contribution aside, it is a burden on the mind of the testee because the test stem asks the testee to choose the right option "according to the picture". The testee will spend his precious time for trying to find a clue in the picture while answering this question. The test writer might have thought that the picture would be helpful in the understanding of the word *birthday*. However, knowing the meaning of the word *birthday* does not help the testee to activate his knowledge about the choice of prepositions. In order to test this hypothesis, we administered this item to 32 seventh graders by dividing them into two groups. 15 of these students received the test with the picture above and the other group of 17 students received it without the picture. It was observed that 9 students out of 15 (60%), who took the item with the picture, marked the option B, which is the right answer. In the group who took the item without the picture, 10 students out of 17 (58.8%) marked the option B. The difference between the two groups is 1.2% on behalf of those who took the item with picture. So, it means that the removal of the picture from this test item does not weaken the communication between the test taker and the test item to a significant degree.

Testing Linguistic Communication or Explicit Linguistic Knowledge

If teachers as test writers want to write valid tests, they must pay utmost attention to the correlation between their language focus in teaching and their focus in testing. If their

teaching is based on the initiation, maintenance and termination of communication, then the focus of the same teacher cannot be the possession of explicit linguistic knowledge by students in testing. Test item (6) is extracted from a test that was administered throughout Turkey.

(6)   Soruda altı çizili verilen sözcüklerden hangileri zarftır? (Which ones of the underlined words in the question are adverbs?)

I. Yasin got very bad marks. He should study hard.

II. New planes are very fast.

III. The little baby is very lovely.

IV. Merve goes to music courses regularly.

A) I-III   B) I-IV   C) II-IV   D) III-IV

It is known that there is a very strong relationship between teaching-learning practices and testing. The validity of this test item depends strongly on whether teacher's aim is to raise the linguistic consciousness in the learner or not. If the teacher aims a maintainable communication by the students in the classroom and follows a communicative syllabus, this test item has problems related to validity. Communicative syllabuses avoid containing such items since this type of items raise grammatical consciousness in learner (Richards and Rogers, 1990:67) which spurts some degree of anxiety of making error that stems from overmonitoring (see Krashen (1987:16). Communicative syllabuses place primary emphasis on the use of target language to communicate in meaningful situations rather than possessing explicit metalinguistic knowledge. The evaluation of the test takers' answers must also be in the way communicative language teaching suggests. For instance, the right option in item (7) is given as B in the key.

(7) Aşağıdaki soruda verilen resme göre konuşma balonunda boş bırakılan yere uygun gelecek ifadeyi seçiniz.

A) What would you like to drink, sir

B) What would you like to eat sir

C) Do you like fish, sir

D) Would you like some green salad, sir

There is nothing wrong with this option's being the right one. However, how would the teacher's attitude be towards the test takers' opting for C? Although the structural features of the answer from the customer urges the test taker to choose the option B, is the option C totally wrong if the waiter utters the question "Do you like fish, sir?" to mean "If you like fish, sir, I am going to recommend a type of fish which is our special." and the customer utters that sentence to mean "Yes I like fish but I don't want only fish. I would like both fish and chips." Is it a far-fetched scenario? If linguistic elements are given life within the context they are used, then this scenario is just one that is extracted from the heart of daily life. Another problem with this item is its dictating power in the setting of the parameters of the target language. Parameters of a language are set through binary options that are mutually exclusive. For instance, an English child acquiring her fist language sets the parameters of prepositional phrase structures in English by inducing the rule from the occurrences of this structure in meaningful situations. Once this hypothetical child is exposed to the phrases such as "in the room, on the table, behind the door, etc." she sets the rule dictating that "English is *head initial* regarding prepositional phrases" as opposed to the choice "English is *head final* regarding prepositional phrases." In test item (7), once the learner is told that the option B is the only one which is right, others, especially option C, which competes with the option B, would be registered as "wrong". Setting such a parameter would confine the language user to the use of "What would you like to eat, sir" structure in such situations, whereas a native speaker of English would enjoy the flexibility of using very different structures to mean what meant with the option B.

If the teacher demands her students to be perfect in the target language, as behavioristic approaches demand, the option to be chosen is obviously B. However, if the teacher is

after a maintainable communication in the target language to foster oral production by students, as communicative and constructivist approaches require, then the option C would also be convenient.

CONCLUSION

The fact that scores of a test is so crucial in the decisions given about an individual's roles in a given community urges test writers to be careful to utmost degree while writing tests. The analyses of some of the items in language tests that were administered locally, city-wide or throughout Turkey reveal that there are five types of deficiencies in test items that stem from the discrepancy between the test writers' assumptions about their audience while writing test items and the test takers' conceptualization of these test items. The first discrepancy stems from the assumed goal of the test item and the way it is consumed by the test taker on the plane of cognition. In this type, a test writer writes a test item to test linguistic skills of a language learner. However, the test item turns out to be invalid or unfair because it tests cognitive skills of test takers rather than or as much as their linguistic skills. The second type of deficiency emerges because test items require topical or general knowledge from test takers rather than testing their comprehension and usage of language in a given situation. The third one is developmental in the sense that test takers are not either mature and experienced enough to possess the knowledge that is being asked or they are not capable of conceptualizing the test item in the way the test writer assumes them to. The fourth deficiency we identified is related to the use of visual aids in test items. It was observed that some test items contain pictures which are not at least functional in fostering the test takers' processing of the test item. The last one is related to the discrepancy between what current language teaching approaches suggest in language learning and the way the test items approach to the question "What does it mean to know a language?". It was observed that some test items seek explicit linguistic knowledge in the mind of test takers rather than seeking how they communicate in the new language they learn.

In order to minimize such deficiencies in test items, if it is not possible to get rid of them totally, a local language test must be administered to a small number of testees as piloting before it is placed before the actual test takers and the results must be evaluated by peer teachers, rather than the test writer himself or herself to overcome the deficiencies in the test item; the items in a nationwide language test must be checked by experts from different fields such as linguistics, psychology, cognition and pedagogy.

REFERENCES

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: OUP

Bachman, L. F. and Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests*. Oxford: OUP.

Baghaei, P. (2011). Optimal number of gaps in C-test passages. *International Educational Studies*, 4(1), 166-171.

Bakhtin, M. M. (1981). *The Dialogic Imagination: Four Essays*. Ed. Michael Holquist. Trans. Caryl Emerson & Michael Holquist. Austin and London: University of Texas Press.

Benedetti, K. D. (2006). Language testing: Some problems and solutions. *MEXTESOL Journal*, 30(1), 25-40.

Brown, J. D. (1996). *Testing in language programs*. New Jersey: Prentice Hall.

Cunning-Wilson, C. (2001). Choosing EFL/ESL visual assessments: image and picture selection on foreign and second language exams. *ERIC*. Document code: ED 452 707.

Currie, M. and Chiramanee, T. (2010). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing*, 27(4), 471-491.

Fazeli, S. H. (2010). The Impact analysis of psychological reliability of population pilot study for selection of particular reliable multi-choice item test in foreign language. *Journal of Language and Linguistic Studies,* 6(2), 7-21.

Fulcher, G. and Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.

Gershkoff-Stowe, L., Thal, D. J., Smith, L. B. and Namy, L. L. (1997). Categorization and its developmental relation to early language. *Child Development*, 68(5), 843-859.

Gershkoff-stowe, L., Connell, B. and Smith, L. (2006). Priming overgeneralizations in two- and four-year-old children. *Journal of Child Language*, 461-486.

Hambrick, D. Z., Sathouse, T. A. and Meinz, E. J. (1999). Predictors of crossword puzzle proficiency and moderations of age-cognition relations. *Journal of Experimental Psychology: General*. 128(2), 131-164.

Heaton, J. B. (1990). *Writing English language tests*. London: Longman.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: CUP.

Ironson, G., Homan, S., Willis, R. and Signer, B. (1984). The validity of item biased techniques with math word problems. *Applied Psychological Measurement,* 8(4), 391-396.

Izard, J. (2005). Overview of test construction. In Kenneth N. Ross. (Series Ed.) *Quantitative research methods in educational planning*. Module 6. Paris: UNESCO International Institute for Educational Planning.

Kane, M. (2010). Validity and fairness. *Language Testing*, 27(2), 177-182.

Krashen, S. (1987). *Principles and practice in second language acquisition*. London: Prentice-Hall International.

Larner, A. (2009). Neuropsychology of board games puzzles and quizzes. *Advances in Clinical Neuroscience and Rehabilitation*, 9(5), p. 42.

Park, C. (2010). Differential item functioning analysis of an EFL vocabulary test. *English Teaching*. 65(3), 23-41.

Perrone, M. (2006). Differential item functioning and item bias: Critical considerations in test fairness. *Columbia University Working Papers in TESOL & Applied Linguistics.* 6(2), 1-3.

Piaget, J. (1929). *The child's conception of the World*. London: Routledge and Kegan Paul Ltd.

Piaget, J. (1930). *The child's conception of physical causality*. London: Kegan Paul.

Pinker, S. (2004). Clarifying the logical problem of language acquisition. *Journal of Child Language*. 31, 949-953.

Richards, J. C. and Rodgers, T. S. (1990). *Approaches and methods in language teaching: A description and analysis*. Cambridge: CUP.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. and Boyes-Bream, P. (1976). Basic objects in natural categories. *Cognitive psychology,*. 8, 382-439.

Subvokiak, M., Mack, J., Ironson, G. and Craig, R. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. *Journal of Educational Measurement, 21*, 49-58.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Hampshire: Palgrave Macmillan.

Withers, G. (2005). Item writing for tests and examinations. In Kenneth N. Ross. (Series Ed.) *Quantitative research methods in educational planning*. Module 5. Paris: UNESCO International Institute for Educational Planning.

Xi, X. (2010). How do we go about investigating test fairness. *Language Testing*, 27(2), 147-170.

UZUN GENİŞ ÖZET

Giriş

Hiç bir sınav sorusu, o soruyu cevaplayacağı varsayılan kişileri hesaba katmadan yazılamaz. Sınav hazırlayanın varsaydığı öğrenci ile gerçek öğrencinin uyuşmazlığı, sınav başarısında bir etken olarak karşımıza çıkmaktadır.

Bu çalışmanın amacı, ülke çapında ya da sınıfta uygulanmış soru örneklerini inceleyerek, sözü edilen uyumsuzluğun neler olduğunu ortaya koymaktır.

Yöntem

Geçmiş yıllarda Türkiye genelinde veya sadece sınıfta uygulanmış yabancı dil soruları taranarak, soru yazarının hedefi ile gerçekleşen durum arasındaki uyuşmazlıklar, nitel bir bakış açısıyla belirlenip değerlendirilmiştir.

Bulgular

Yapılan inceleme, sözü edilen uyuşmazlıkların, aşağıda verilen beş nedenden dolayı ortaya çıktığını göstermiştir.

Dil mi, Bilişsel Beceriler mi?

Bir sorunun geçerliği, sorunun yalnızca ölçmek istediği bilgi veya beceriyi ölçmesi, bunun dışında herhangi bir ölçüm yapmamasına bağlıdır. İncelenen bazı yabancı dil sorularının, öğrencilerin dil becerilerinden fazla, bilişsel becerilerini ölçtüğü görülmüştür.

Alan Bilgisi mi, Dil mi?

Bazı soruların cevaplanabilmesi için alan bilgisi bilmek, dili kullanabilmenin çok önüne geçmektedir. Örneğin, ülkelerin başkentlerini soran bir soru, öğrencinin hatırlama

becerileriyle, genel dünya bilgisini ölçmektedir. Öğrencinin, Atina ve Yunanistan'ı gördüğünde, bu ikisini eşlemek için, dil bilmesinden fazla genel bilgiye gereksinimi vardır.

Sınava Girenlerin Yaşı ve İçerik Uyuşmazlığı

Sınav soruları hazırlanırken, öğrencilerin yaşları mutlaka dikkate alınmalıdır. Bulunduğu ilin dışına pek çıkmamış bir öğrenci için, Türkiye'yi temsil edecek bir yiyecek, o ildeki en meşhur yiyecektir. Ülkelerle, o ülkelere ait yiyeceklerin eşlenmesinin istendiği bir soruda, bazı 4. sınıf öğrencilerinin eşlemede başarısız olduğu, bir öğrencinin de sorunun altına, "Turkiye için 'ceviz ezmesi' olması lazım ama o da şıklarda yok" diye not düştüğü, yani, küçük yaşlardaki öğrencilerin bilgileri aşırı genelleştirdiği, bu nedenle, soruyu doğru cevaplayamadığı gözlenmiştir.

Sorulardaki Görsel Ögelerin İşlevi

Görsel öge içeren dil sorularının, görsel kısımları sorunun ayrılmaz bir parçasını oluşturmalı, görsel öge ya sorunun anlaşılmasını kolaylaştırmalı ya da soru o görsel öge olmadan cevaplanamamalıdır. İncelenen sorulardaki bazı görsel ögelerin soruyu cevaplayana hiç katkısının olmadığı deneysel olarak gösterilmiştir.

Dilsel İletişim mi, Yoksa Dilbilgisi mi?

Günümüz dil öğretim kuram ve yöntemleri, yabancı dilin sınıf içinde gerçek gereksinimleri karşılamak için kullandırılmasını önerir. Oysa bazı sorular, açık dilbilgisi sormaktadır. Öğrencilerden dilbilgisi kuralları isteyen dil soruları öğretim süreçlerini de o şekilde yönlendireceğinden, bu tür sorular dil öğretimi için zararlıdır.

Sonuç

Bu çalışmada, yabancı dil sınavlarını hazırlayanların beklentileri ile sınava giren adayların sorulara yaklaşımları arasındaki uyuşmazlık ele alınmıştır. Çalışmanın bulguları, uyuşmazlığın beş nedenden kaynaklandığını ortaya koymuştur. Bazı dil sorularının, öğrencilerin dil becerilerinden çok bilişsel becerilerini ölçtüğü, bazı soruların ise dil becerisiyle değil, genel bilgi bilmekle çözülebildiği gözlenmiştir. Sınav

hazırlayanın, sınava girecek adayların yaşlarını, dolayısıyla yaşam deneyimlerini de dikkate alması gerekmektedir; bazı öğrencilerin bilgiyi aşırı genelleştirdiklerinden dolayı, soruları doğru cevaplayamadığı görülmüştür. İncelenen soruların bazılarında verilen görsel ögelerin öğrencilere hiç bir şekilde yararlı olmadıkları gözlenmiştir. Soruların bazılarının öğrencilerden açık dilbilgisi istediği görülmüştür. Bu bilgi, dilin iletişim aracı olarak kullanılmasında hiç bir işe yaramayacağı için, açıkça dilbilgisi soran soruların, günümüz yabancı dil öğretim kuram ve yöntemleri dikkate alındığında, geçerliği yoktur.

Kısaca, ülke çapında uygulanan sınavları hazırlayanlar, hazırladıkları soruları, psikolog, eğitimci, dil bilimci ve benzeri uzmanların incelemesinden geçirmelidirler. Sınıf içinde uygulanacak sınavlar ise, tek bir öğretmen tarafından değil, mutlaka bir sınav takımı tarafından hazırlanmalıdır.