Education and Science

Vol 43 (2018) No 194 261-284

Computerized Adaptive Testing Design for Students with Visual Impairment *

Selma Şenel [1], Ömer Kutlu [2]

Abstract

Computerized Adaptive Testing (CAT), a product of scientific advancements and new technologies in the field of measurement and evaluation in education, may be considered beneficial in reaching more valid and reliable test results for students with visual impairment. CAT renders more reliable test results in less time with fewer items that are well-matched to the ability levels of students. When these basic properties of CAT are considered, we can state that the use of CAT in assessing the ability levels of students with visual impairment may provide a significant step in attaining fair and comparative test results. The aim of this research is to develop audio-based CAT software for visually impaired students and to guide future studies by providing assessment, development, and design procedures. We created an item pool comprised of 166 audible items to assess the listening comprehension levels of students with visual impairment at secondary school. We have reported all data that indicate the validity and reliability scores through test development process, such as the acquisition analyses, expert opinions, pre-trial application (n=196), trial application (n=608), assumptions for the Item Response Theory. We have depicted of all the steps of software development on the basis of multimedia design principles in detail. Parameter estimates and software development studies were made according to 3 parameter logistic model. During item development and software testing, opinions of 7 students with visual impairment were considered. As a result, a valid and reliable CAT, named *SesliTest*, has been developed for students with visual impairment at secondary school. The research carries great importance since it proves CAT is applicable for students with visual impairment.

Keywords

Computerized adaptive testing
Test accommodation
Visually impaired students
Listening comprehension
Students with special needs

Article Info

Received: 09.19.2017
Accepted: 03.14.2018
Online Published: 04.10.2018

DOI: 10.15390/EB.2018.7515

* This article is derived from Selma Şenel's PhD dissertation entitled "Investigation of the compatibility of computerized adaptive testing on students with visually impaired", conducted under the supervision of Ömer Kutlu.

[1] Balıkesir University, Information Processing Application and Research Center, Turkey, selmahocuk@gmail.com
[2] Ankara University, Faculty of Education, Educational Measurement and Evaluation, Turkey, omerkutlu@ankara.edu.tr

Introduction

Many critical decisions; such as passing a course, going to a higher education institution and also getting a job are depend on the results of achievement tests. The most important issue in these tests is the validity of the test results (Haladyna & Downing, 2011). The main condition for the validity of tests is to prevent the variables, other than the variables wanted to be assessed, from affecting the test score. Test scores of special needs students may include threats to validity because their specific requirements or disabilities may affect test scores directly. For example it is meaningless for a visually impaired student to take visual tests. To fulfil validity, ability levels of special needs students must be defined free from the effects of their disabilities. In order to eliminate the effects that may arise because of the status of special needs students, the test material, or the testing process, can be adjusted to students' needs, without causing any change in the measurement of the construct. Test accommodations are changes that are intentionally made for special needs students on learning environments, testing materials or testing conditions. (National Center for Learning Disabilities [NCLD], 2005; Tindal, 1998). It is assumed that the testing conditions for the disabled students and other students are equalised with these accommodations (Almond, Lehr, Thurlow, & Quenemoen, 2002). Extending test duration, increasing font sizes used in the test booklets, or providing expert assistance, such as a human or computer reader, are among the widely used accommodations (Educational Test Service [ETS], 2014; Ministry of National Education [MEB], 2012, 2013; Center for Evaluation, Selection and Placement [ÖSYM], 2013; Stone & Davey, 2011).

Meanwhile, test accommodations may create some other validity threats. Appropriateness of the test accommodation to the specific disabled group, the application process, and the characteristics of the construct wanted to be assessed with the testing are debating issues in the literature (American Educational Research Association [AERA], American Psychological Association [APA], & National Council of Measurement in Education [NCME], 1998; Bielinski, Thurlow, Ysseldyke, Freidebach, & Freidebach, 2001; Clapper, Morse, Thompson, & Thurlow, 2005; Higgins & Katz, 2013; Koretz & Barton, 2003; Sireci, Li, & Scarpati, 2003). One of the major concerns is that the provided testing accommodations may cause changes in the features of the items and of the test itself. Receding from the construct, targeted to be measured, because of these accommodations, will change the psychometric characteristics of the test and the meaning of the test scores (Johnstone, Altman, & Thurlow, 2006). Any change that may occur in the test scores will cause problems in comparing the scores of the students.

Another problem is that the students with special needs remain at low proficiency levels at the exams, which are designed according to mid-range ability scale where most of the students accumulate –a situation which causes these students to prefer just pseudo-guess the answers of the questions (Abedi, Leon, & Kao, 2007; Laitusis, Cook, Buzick, & Stone, 2011; Minnema, Thurlow, Bielinski, & Scott, 2000). If the rate of students with disabilities answering the questions by just guessing is increasing, then the results of these assessments will not provide adequate information about the proficiency level of these groups with special needs. Since these students cannot display their real level of proficiency, this may cause anxiety in them and they may start participate less in the tests (Stone & Davey, 2011).

Technological and scientific developments are taking place in measurement and evaluation field in order to obtain more valid and reliable test results for students who remain at each end of the ability distribution. Item Response Theory (IRT) and Computerized Adaptive Tests (CAT) –one of IRT's main applications in psychological measurement- can be reported as the most prominent of all these developments (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). Adaptive tests are used according to students' ability levels. In general terms, in adaptive tests, the next item is determined according to the performance the student had showed in the previous item (Hendrickson, 2007). Although there are different algorithms, simplest CAT tries to determine the ability level of examinee by sending easier items if examinee answers incorrectly or CAT sends harder items as the examinee replies correctly (Accountability and Curriculum Reform Effort [ACRE], 2010). Traditional tests use predetermined items that are chosen for moderate ability levels. While examinees are administered a fixed test-form containing pre-determined multiple-choice questions in traditional tests; the items and

the number of items can be different for each examinee, in adaptive tests since adaptive tests uses items depending on students' performance (Raiche & Blais, 2002). As the test items are presented considering student ability, it is possible to make more precise and reliable measurement in a shorter time, with less number of items compared to traditional tests (Hambleton et al., 1991; Kingsbury & Hauser, 2004; Rudner, 1998; Thompson, 2010; Thurlow, Lazarus, Albus, & Hodgson, 2010; Tian, Miao, Zhu, & Gong, 2007; Wainer, 2000). Hereby, guessing rates will decrease since students will reply items that are close to their ability level. Besides, even though they replied different items and test forms, total student scores can be compared to each other by IRT's invariance feature (Bennett, 1999; Stone & Davey, 2011; Wainer, 2000). According to the invariance feature of IRT; the ability parameter is independent from item sample (Hambleton et al., 1991). This would allow fair comparison of the scores of special needs students who tend to be at lower ability levels, with their peers.

The scientific works and developments, and the technological advancements generally focus on making life more comfortable and independent for visually impaired students (Erdem, 2017; Papadopoulos & Goudiras, 2005). In today's world, where new technologies fast emerge from the purpose of making human life easier, and the assessment and evaluation applications are becoming more and more important in shaping people's lives, it is as important to fully benefit from the technological advancements in the tests administered to the students with special needs. The advancing technologies and scientific developments, even before facilitating a better life for students with normal developments, must help out the students with special needs to overcome the challenges and difficulties they face with in their lives. Thus, a fairer educational environment can be provided.

After the computer technologies have become nearly inseparable part of their daily lives, these technologies started to be used also in the tests especially adapted for students with disabilities. Thanks to computer technologies, the students with disabilities can carry out their life activities such as working, studying, communicating, learning and entertaining, more independently. For example, students who have difficulty in using pencils can continue their writing with word-processors; those who cannot read, can listen to the texts with synthesizers; visually impaired people can read the printed or screened texts transcribed in Braille; physically disabled people can engage in social life trough virtual environments (Bennett, 1999; Laitusis, Buzick, Stone, Hansen, & Hakkinen, 2012; Yaman, Dönmez, Avcı, & Kabakçı Yurdakul, 2016).

It has been observed that specially the visually impaired students see themselves capable of using computer technologies confidently and are very keen to use computer technologies in the tests they are administered (Higgins & Katz, 2013; Karabay, 2016; Şenel, 2015; Tavşancıl, Uluman, & Furat, 2012). Despite their eagerness, the widest accommodation where these technologies are applied for visually impaired students is only in form of a human or virtual loud reader (Anastasi, 1988). Visually impaired students, who have used loud readers, state that their success in the tests depends on the reader's quality (Sánchez & Espinoza, 2012; Karabay, 2016). Visually disabled students think that the socio-cultural behaviours, instructional qualifications, pronunciation and reading skills, reading speed, and the expertise field of the readers do affect their performance at the tests. For instance, a reader who is specialised in a different field than the responder, who can have difficulty in pronouncing the special English words correctly, or a sleepy reader reading the text in a bored manner, diversely affect the performance of the visually impaired examinees. Meanwhile, the visually impaired examinees, because they are not able to do the markings themselves, can suffer from some psychological conditions like uncertainty, diffidence, anxiety, feeling embarrassed to ask the reader to read the question again or from giving an incorrect answer, or loss of attention (Abell & Lewis, 2005; Şenel, 2015).

It is clear that even the most basic features of CAT can offer many improvements for testing the visually impaired students. The most significant of them is obtaining more reliable and valid test results for the visually impaired students by administering them the tests that are more appropriate to their ability levels. Tests with shorter durations, but providing more reliable test results and better information about the students' level of ability, can help raise the examinees' confidence and increase the rate of participation in the tests (Clark, 2004; Thompson, 2010). Moreover in CAT, it is possible to

ask questions, difficulty level of which is lower than the students' ability level, to keep their motivation and confidence high (Hausler & Sommer, 2008). Apart from motivation and confidence, CAT can also provide some other positive psychological effects. For example; CAT facilitates the use of computer-based loud readers instead of a human reader, which helps visually impaired students to feel less embarassed during the testing (Abell & Lewis, 2005; Stone & Davey, 2011; Şenel, 2015). On the other hand, because the CATs take a short time to complete the test, it can be a good alternative to the "extended testing duration" accommodation, given for all the students with special needs. This way, the boredom, fatigue and loss of attention, caused by the extended-testing, can be eliminated. Another improvement is about the physical accommodation. As the testing is held in a computerized environment, the examinees are able to modify colour contrasts, use loud reader as they wish, make changes in visual presentations, interpret sign language, increase text fonts, and/or filter the contents. Moreover in CAT, many accommodations can be supplied in a much cheaper way than the classical paper-pencil tests (Russell, Higgins, & Hoffmann, 2009).

The aim of this study is to design and develop a CAT application for secondary school visually impaired students that will help determine their ability level of listening comprehension. In the literature, we did not come across any CAT application developed specially for visually impaired students. Visually impaired students pay more attention to sounds and to listening, to be able to understand surroundings. In other words, their most important "receptive language" is listening. Because of this, the ability of understanding what they listen is very important to give meaning to their lives. Taking this importance into consideration, the ability of comprehensive listening is defined as the foundation of this study. In this paper, we present the results of our study, which aimed to develop CAT software using an item pool, through which the visually impaired students will be able to demonstrate their listening comprehension skills.

Method

The Study Group

To examine the assumptions of the Item Response Theory and to define the item parameters, we need large-scale sample cases (Hambleton, 1990). Since we have not been able to meet this requirement with only the visually impaired students to whom we could reach, we have conducted the trial testing on students who have no visual disability. The feature of the IRT that provides "estimation for the item parameters independently from the ability level of the responder group", has allowed us to apply the trial tests on groups other than the group targeted for the application. Under this assumption, we have chosen a study group of 608 students from the secondary level pupils studying their 2016-2016 academic year, at four schools of *Balıkesir* situated at the city centre and in central provinces. Boys comprised the 47.37 %, and girls comprised the 52.63 % of the study group. Comprising the 56.09 % of the study group, the 8th graders were more in number than the 7th graders.

Assumptions and Limitations

- The trial application data that defined the item parameters are limited to 608 students.
- The trial application has been conducted in eight sessions that are held at different times. It is assumed that the measured quality did not change within these time periods.
- It is assumed that the item parameters obtained by applying the tests on students with no disabilities carry the same values as if they would have been applied to students with visual impairment.

Creation of Listening Comprehension Item Pool

To be able to apply CAT in our assessment, we needed first to create an item pool, composed of quality items that can address a wide range of ability levels. In this section, we explained the steps we had taken to create our item pool.

Analysis of Acquirement and Intellectual Level

Since as part of our study, we were going to develop a test that is adapted to student needs, the requirements, such as the need of estimating the ability level of the responder straight after they answer an item, and keeping the test score independent from the responder's language skills, have limited the types of items we could use in our test. In the study, we used multiple-choice items, because they provide have a wide usage area, and provide ease of application and scoring.

In primary school level Turkish language lessons, the field of listening comprehension is categorised in five sub-fields as follows: (a) Applying the listening rules, (b) Understanding and analysing the listened topics, (c) Evaluating the listened topics, (d) Enriching the lexicon, and (e) Acquiring an effective listening habit (MEB, 2006b). In line with the thoughts of an academic, who is expert in measurement and evaluation and has got works in writing Turkish test items, we decided that, since they are from the acquirements that cannot be measured with multiple-choice items, the sub-fields of "applying the listening rules" and "acquiring an effective listening habit" should be left out of our study. Thus, 14 out of the 42 listening comprehension items (MEB, 2006b) have been left out of the study.

The Listening Comprehension Ability Test has been prepared based on the remaining 28 acquisitions. Because they are based on an internationally accepted classification that is directly related to comprehensive skills, these acquisitions have been grouped according to the four comprehension stages that are used in the Progress in International Reading Literacy Study (PIRLS) (MEB, 2003):

1st Stage: Retrieving clearly stated thoughts in the text (direct deduction)

2nd Stage: Finding and inferring the not-very-clearly expressed thoughts in the text (direct deduction)

3rd Stage: Linking the events explained in the text with personal information and experiences (integrating and interpreting the ideas and information)

4th Stage: Examining and evaluating the elements, content, and language of the text

In our study, we matched the comprehension processes of PIRLS with the scope of our test and the behaviours that will be assessed. These matchings have been administered under the suggestions of an academic who is specialised in the field of assessment and evaluation. The matching of the sub-fields and acquisitions of the test with the expression unit, intellectual levels and number of the developed items are given in the Addendum 1.

Developing Test Items for Listening Comprehension

Considering the difficulty of fulfilling the assumption of wide-range item pool, comprised of quality CAT items in developing the items; initially we wanted to examine the items that had been developed by an expert team within a plan. For this purpose, we examined the Turkish items developed for the National Secondary Schools Student Selection and Placement Exam (*OKÖSYS*). *OKÖSYS* had been developed and administered by the Ministry of Education between years 1998 and 2007 to select students for secondary education. *OKÖSYS* tests aimed to measure upper level intellectual processes such as using, interpreting, generalising, estimating, breaking down into components, establishing correlation between components, and evaluating the existing knowledge (Kutlu & Karakaya, 2004). Our examination of the items has showed that items related to reading comprehension comprise the majority. Judging that abilities of reading and listening are the two receptive languages that are very similar to each other (Özbay, 2005) and, as also stated in the literature, the developments of reading comprehension and listening comprehension are mutually support each other (Kutlu, Bilican, & Yıldırım, 2010), we have used the items, prepared for reading comprehension. Some items that are defined as corresponding to the acquisitions, on which the listening comprehension test was based, were directly selected, and the texts that can be used for developing new items, have been determined. In this way, we have used 93 *OKÖSYS* items. To determine the listening texts, we also examined the items in the field of Turkish, asked at Student Selection Exam –a national exam developed and administered by the Turkish Student Selection and Placement Centre to admit students for universities– between 1999 and 2010. As a result of this work, we selected nine reading texts that are suitable for 7th and 8th graders and used them in our test as listening texts.

We have taken an extra care to provide texts, the level, length, and vocabulary –both in variety and in number– of which are appropriate to the level of the students. Since in daily life, students are exposed to texts written in different genres, we also paid attention to choosing a variety of texts for our test items, selected from different genres, with different contents, used for different purposes (personal, public, educational or occupational). Short stories, poems, narrations, articles, diaries, critical reviews, editorials, essays, and conversations are from the genres we included in the test. The item development process has been carried out in company of a Turkish teacher. After the initial development stage, we received the expert opinions of three academics and four PhD students who are specialised in the field of assessment and evaluation.

As a result of the adjustments, conducted in line with the suggestions of the experts, we initially developed 232 items. These items were arranged in seven, 33 and 34-item test forms to be used for preliminary testing. In the 2014-2015 academic year, we administered the preliminary tests to 196 students –100 from the 7th graders and 96 from the 8th graders. The data obtained from the preliminary tests have been analysed in Microsoft Excel in accordance with the Classical Testing Theory and reported as shown in the sample item card in Figure 1. This report was presented to the review of an academic who is experienced in test development processes and has works in writing Turkish test items and is proficient in the field of assessment and evaluation.

As a result of the preliminary testing, among the 232 developed items, 26 out of 44 items –which were under the 0.20 per cent item discrimination index– were left out of the study as recommended by the expert. The remaining 17 items were edited under the observation of the expert and added back into the test application item pool. For the reason that the number of items has reduced during some learning acquirements, two items from the Secondary Education Private Schools Entry Exams (*ÖOS*) held in 2002 and 2003, have been added to the pool as new items. *ÖOS* is an exam developed and administered by Turkish Ministry of Education to select and place students from primary education to private secondary schools, which measure the same abilities as the *OKÖSYS* tests.

Acquisition	A1. Can derive meanings of the words and word-groups from the context of listened material.
Intellectual Level	1st Stage: Inferring the clearly expressed thoughts in the text

Item No:	10101	**Evaluation**
Expression Unit	Word	Discrimination: Very good, but can be developed further Difficulty level: Very easy Distractors: We can work more upon A and C distractors
Text Resource	OKS 2000	
Text Type	-	
Correct Choice *	B	

A	B*	C	D	Unmarked	P_j	q_j	S_j	α_3	α_4	r_{pbis}	$r_{\%27}$	r_j
1	28	1	7	0	0,74	0,26	0,44	-1,1	-0,8	0,31	0,40	0,14

"Başından geçenleri sakin sakin anlattı." cümlesinde "sakin sakin" ikilemesi cümleye aşağıdaki anlamlardan hangisini vermek için kullanılmıştır?
A) Yaşanılanların çok önemli olmadığını dile getirmek
B) Davranışlarda bir heyecan bulunmadığını anlatmak
C) Anlatan kişinin yaşadığı üzüntüyü ifade etmek
D) Konuşulanların bu biçimde daha güzel anlaşılacağını vurgulamak

Figure 1. Sample Item Card

From the 208 items in our item pool, we created eight 26-item forms. Since we observed in the trial test that the students were having difficulty to complete answering one 33-34 item test form in one course hour, we edited this trial form to contain 26 items.

The trial test forms have been administered by loud-reading to four students with visual impairment –two of whom studying at the 7th grade and two studying at the 8th grade– as two forms per student. The test items have been read out-loud one-by-one and the reviews of the students about the items have been taken. We observed that these students had difficulty in answering the test items that require completing blanks in sentences, and the items that have underlined expressions. Under the light of the information we obtained from both our observations during the trial process and the negotiations we held with the experts, we edited these items. An item, on which we could not make any change, has been removed from the pool, thus the number of items decreased to 207. Appropriateness of the items have been checked by two Turkish teachers –one of whom is visually impaired– who are serving at secondary schools for students with visual impairment. The teachers approved that the items are appropriate to be administered to the students with visual impairment. The table of specifications related to the items took shape after the final editions, and the number of items according to the expression unit are given in Table 1 and Table 2.

Table 1. Table of Specifications

	Intellectual Level				
Sub-Field	**1. Stage**	**2. Stage**	**3. Stage**	**4. Stage**	**Total**
Understanding and analysing the listened topics	40	69	58	-	167
Evaluating the listened topics	-	-	-	14	14
Enriching the lexicon	-	8	18	-	26
Total	40	77	76	14	207

Table 2. Anlatım Birimine Göre Madde Sayıları

	Intellectual Level				
Expression Unit	**1. Stage**	**2. Stage**	**3. Stage**	**4. Stage**	**Total**
Paragraph	34	54	52	14	154
Word	6	8	7	-	21
Sentence	-	15	17	-	32
Total	40	77	76	14	207

After one item is removed from the pool, eight test forms –one of which consists of 25 items and seven of which consist of 26 items– have been made ready for the trial application. The test forms were designed to provide equivalence in acquisition and choice distribution.

The trial application has been conducted in one month, with eight separate forms, containing 207 items in total. To be able to minimise the incorrect answers given because of lengthy texts, we provided the students answer the forms in eight sessions. The trial test has been administered to 664 responders in total. Since there were numerous sessions, some students could not answer some of the forms due to absence or being in charge of a duty. To reduce any likely data loss, we re-administered test forms to students who were short of two forms. We kept 47 students, who did not answer more than 20 per cent of the items –or two forms–, out of the study. We also excluded 9 students, who selected the same answers in all the forms or who answered completely randomly, from our study. In conclusion, the analyses have been conducted over 608 students. The lost values in 608-person data have been covered with the mean value assignment method (Mertler & Vannatta, 2005).

Results

Checking the Assumptions of Item Response Theory

In this section, we examined whether the 207-item trial test, administered to 608 students, have confirmed IRT or not.

Unidimensionality

In order to find out whether a measured property has a unidimensional structure or not, we carried out an exploratory factor analysis (EFA). Since the obtained data had two (binary) categories, we first created a tetrachoric correlation matrix (Embretson & Reise, 2000, p. 37). We interpreted obtaining a dominant factor at the end of the EFA we ran in Mplus software program using the tetrachoric correlation matrix, as the signal of unidimensionality (Crocker & Algina, 1986; Hambleton et al., 1991). The eigenvalues about the components we obtained as a result of the EFA have been examined.

It is seen that the eigenvalue of the first factor is 45,520, which is too much above 5,879, the eigenvalue of the second factor. This may suggest that only one factor is dominant, and the rest of the factors are nonsensical. In view of this data, we decided that the test is unidimensional. At the end of the EFA, we examined the factor loadings of the items under unidimension and removed 35 items, factor loadings of which are below 0.30, from the items pool. After the extraction of these items, it is seen that this extraction did not make such a big change that can create a great difference in the acquisition distribution in the items pool, hence cause a negative effect on the measurement of the comprehension ability. The factor loadings of the remaining 172 items are given in Table 3. When we examined the table, we saw that factor loadings of 85 items are over 0.50, which explain very well the variation of the measured property of the items in the pool.

Table 3. Factor Loadings of the Items

Item No	Factor Loading	Item No	Factor Loading	Item No	Factor Loading	Item No	Factor Loading	Item No	Factor Loading
M1	0.298	M36	0.630	M71	0.685	M106	0.669	M141	0.533
M2	0.353	M37	0.433	M72	0.560	M107	0.414	M142	0.712
M3	0.446	M38	0.474	M73	0.532	M108	0.436	M143	0.492
M4	0.347	M39	0.515	M74	0.614	M109	0.637	M144	0.516
M5	0.445	M40	0.475	M75	0.553	M110	0.449	M145	0.486
M6	0.488	M41	0.380	M76	0.634	M111	0.606	M146	0.578
M7	0.394	M42	0.475	M77	0.626	M112	0.522	M147	0.337
M8	0.417	M43	0.544	M78	0.464	M113	0.606	M148	0.461
M9	0.413	M44	0.323	M79	0.399	M114	0.701	M149	0.469
M10	0.449	M45	0.538	M80	0.342	M115	0.673	M150	0.301
M11	0.339	M46	0.501	M81	0.501	M116	0.441	M151	0.595
M12	0.503	M47	0.663	M82	0.612	M117	0.657	M152	0.581
M13	0.449	M48	0.595	M83	0.579	M118	0.637	M153	0.539
M14	0.422	M49	0.450	M84	0.559	M119	0.525	M154	0.385
M15	0.443	M50	0.401	M85	0.389	M120	0.559	M155	0.391
M16	0.445	M51	0.640	M86	0.644	M121	0.456	M156	0.553
M17	0.406	M52	0.453	M87	0.348	M122	0.632	M157	0.462
M18	0.348	M53	0.408	M88	0.517	M123	0.463	M158	0.466
M19	0.415	M54	0.651	M89	0.427	M124	0.330	M159	0.656
M20	0.501	M55	0.624	M90	0.437	M125	0.576	M160	0.487

Table 3. Continued

Item No	Factor Loading	Item No	Factor Loading	Item No	Factor Loading	Item No	Factor Loading	Item No	Factor Loading
M21	0.559	M56	0.582	M91	0.667	M126	0.373	M161	0.474
M22	0.376	M57	0.651	M92	0.581	M127	0.374	M162	0.329
M23	0.600	M58	0.439	M93	0.433	M128	0.344	M163	0.599
M24	0.365	M59	0.540	M94	0.599	M129	0.451	M164	0.458
M25	0.458	M60	0.451	M95	0.300	M130	0.527	M165	0.486
M26	0.514	M61	0.499	M96	0.436	M131	0.438	M166	0.304
M27	0.649	M62	0.474	M97	0.427	M132	0.462	M167	0.307
M28	0.491	M63	0.651	M98	0.627	M133	0.369	M168	0.573
M29	0.628	M64	0.614	M99	0.614	M134	0.556	M169	0.587
M30	0.520	M65	0.349	M100	0.519	M135	0.599	M170	0.471
M31	0.500	M66	0.413	M101	0.594	M136	0.577	M171	0.381
M32	0.636	M67	0.657	M102	0.564	M137	0.423	M172	0.384
M33	0.620	M68	0.573	M103	0.443	M138	0.601		
M34	0.644	M69	0.524	M104	0.412	M139	0.532		
M35	0.454	M70	0.524	M105	0.633	M140	0.321		

Local Independence

Local Independence means that the answers given to the test items are independent from each other. This property requires that answering these items must be emanating only from the responder's ability that is wanted to be measured. For this reason, it can be stated that, if the test is unidimensional, the local independence assumption is also approved. We interpreted that, as the unidimensionality assumption is verified in this study, the local independence assumption is also verified (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Hambleton et al., 1991).

Model-Data Fit

During the process of examining the verification of the model-data fit assumption, we analysed the whole data and each item, to see which model they best fit. In this regard, we examined the -2 Log Likelihood ratios which shows the data's angle of separation from the model. When we looked at the -2 Log Likelihood ratios, we saw that 3 Parameter Logistic Model (3PLM) had the lowest -2 Log Likelihood ratio –for 1 Parameter Logistic Model, the ratio of -2 Log Likelihood was 121184.5864; for 2 Parameter Logistic Model, the ratio of -2 Log Likelihood was 120018.5663; and for 3PLM, the ratio of -2 Log Likelihood was 119418.7425– and since we observed that there was a meaningful difference between the model and data fitting, we decided that the 3PLM is the best fitting model to our data set (Hambleton et al., 1991).

The examination of the chi-square values revealed that for 1PLM, 90 items are fitting to the model; for 2PLM, 151 items are fitting to the model; while for 3PLM, 166 items are fitting to the model. It is seen that the items, too, fit best to the 3PLM. Because the responders had the chance of being successful by giving random answers due to the multiple-choice structure of the items, the fitting of the data to 3PLM was an expected outcome (Crocker & Algina, 1986; Hambleton et al., 1991). Since it is found that 3PLM is the best fit to the data, we used 3PLM as the base for developing the software and defining the item parameters.

Invariance of Item Parameters

To examine the invariance of item parameters, we divided our data set into two groups, composed of randomly selected 304 responders, and we estimated the item parameters for each group. Then we compared the linear relationship between the item parameters obtained from the first group

and those obtained from the second group with Pearson's Product Moment Correlation Coefficient (Pearson's *r*). The invariance of the items' parameters is verified as this correlation gets higher. When we examined the relationship between the item parameters, we found that the relationship between parameters b is very high ($r = 0.94$; $p < 0.01$), while the relationship between parameters b is at average ($r = 0.56$; $p < 0.01$). The relationship between the discrimination parameters being lower than the relationship between the difficulty parameters of the items is explained as parameter *a* is affected more from the normality of the point distribution (Kelecioğlu, 2001; Kezer, 2013; Özbaşı, 2014). We saw that while the relationships between the parameters a and b are at acceptable rates, the relationship between the parameters c is quite low ($r = 0.28$, $p < 0.01$). In the literature, it is said that this situation is encountered especially in obtaining the invariant property of 3PLM (Özbaşı, 2014; Doğan, 2002; Yıldırım, Çömlekoğlu, & Berberoğlu, 2003). Consequently, we can say that the invariance of the item parameters are obtained at an acceptable level.

Invariance of Ability Parameters

To examine the independence of the ability parameters from the items, we divided the 166 items that fit to the model into two different randomly selected 86-item sets. We re-estimated the ability parameters for each item set and examined the relationship between the estimated ability parameters. We also examined the relationship between the ability parameters estimated from the two item sets and the ability parameters obtained from the whole of the items by using Pearson's *r*. As a result of this, we found that there is a positive, high and meaningful relationship between the ability parameters estimated from the different item sets ($r = 0.91$; $p < 0.01$). While we found that the correlations between the whole of the items and the ability parameters estimated from the two item sets are 0.96 and 0.97 ($p < 0.01$). These high correlation values suggest that the ability estimations belong to the responders are independent from the items set. Under the light of this information, we can say that the 166-item pool confirms the assumptions of the IRT. All of the items in the pool has been voiced by the researcher and the recordings have been saved into the software database, together with the item parameters.

Audio-Based CAT Software

The CAT software for the study has been created by the researcher as an audio-based application with Microsoft Visual Studio 2012. For reasons, such as not needing installation and allowing an easy access, we designed it as an online app. During the process of developing the software, we received support from a software engineer and an instructional designer. To offer our testing online, we bought a domain under the title of "www.seslitest.com" (*SesliTest*), which will remain open for two years. For the database, we have used Microsoft SQL Server 2012 database management system. In this section, we explained the process of developing the software and the algorithm created for the CAT app.

Developing Process

At the stage of developing *SesliTest*, since our targeted audience are visually impaired students, we left *multimedia, spatial contiguity, temporal contiguity, image and modality* principles from the 12 principles proposed by Mayer's (2005) to design instructional multimedia presentations, out of our study. The remaining principles of *coherence, signalling, redundancy, personalisation, voice, pre-training, and segmenting* have provided guidance for the design of our app. The images of *SesliTest* are given in Addendum-2.

We obtained opinions and comments of an instructional designer for every stage of designing *SesliTest*. The arrangements we have conducted in line with Mayer's (2005) principles and the suggestions of the instructional designer can be summarised as follows:

Coherence Principle: This principle requires excluding extraneous elements – audio, text, or images– that are not directly related to the subject, out of the multimedia design. Pursuant to this suggestion, we preferred a plain design in our audio-based application and kept all non-functional sounds out of our design. Considering that students may experience difficulty, we planned a test attendant would carry out the procedures of registration and signing in to the system for the student.

Signalling Principle: In line with the signalling principle which suggests drawing attention to the words regarded as important for the multimedia presentation that is being designed, we have taken extra care in giving voice to emphases and meaningful intonations. We used the audio files editor and recorder Audacity 2.0.6 for the voiced items, and we edited and filed all sound recordings for the items, listening texts, and answering choices separately. We raised our voice to emphasise the expressions wanted to be intensified such as negative meanings in the items.

Considering that a very long voiced instruction given on the main page of *SesliTest* may adversely affect the attention and motivation of the students, we kept the instruction very brief. The main instruction, describing the purpose and scope of the study, the contents and duration of the test, and the main points of the CAT application, has been given to the students orally before they sat for the test.

Redundancy Principle: In the data processing system there are two different channels for visual and verbal processing, capacities of which are quite limited. This principle suggests that learning process can be more successful when the words and graphics are used together, as both audio and text files. However, loading too many elements in the same channel, which has a limited capacity, will take up too much space and eventually hinder good learning (Mayer, 2005). Although the app had text files that are adjusted to big fonts ready to be presented to students with very limited sight, pursuant to the redundancy principle, the items have been presented only as audio files. Moreover, being concerned that the students with limited sight, who can proceed the test by reading the instructions, will have precedence over the totally blind students, we have given space for written texts as few as possible in all the pages of the online media.

Personalisation Principle: This principle suggests that the learning can be better achieved when a daily, conversational style language is used in the media, rather than a formal, academic language. We have designed the instruction in accordance with this principle.

Voice Principle: This principle states that the learning can be more successful when verbal statements are presented with a human voice, rather than a machine voice. In conformance to this principle, all the test items and all of the instructions and texts in *SesliTest* have been voiced over by the researcher.

Pre-training Principle: In pursuant to this principle, we prepared a trial version for the app, through which the students can learn; how the items will be administered, how they are going to listen to the items, how they can replay or rewind, and how they can submit their answers. In our application, in this regard, all exercises that will help students answer the items, are available.

Segmenting Principle: In conformance to this principle, the audio editing has been conducted to ensure that each of the listening texts, question, and answering choices are separately segmented. All of the instructions are given after allowing the responder to complete their task (e.g. listening to the item, entering the answer, submitting the answer, passing to the next item).

CAT Algorithm

Although the methods that can be implemented in a CAT algorithm can differ, there are five main rules to be followed to design a basic CAT algorithm (Weiss & Kingsbury, 1984).

(1) First of all the IRT model to be adopted is determined.

(2) Then an item pool (calibrated item bank) that is appropriate to the model, is created,

(3) The rule of starting a test,

(4) The rule of selecting items,

(5) The rule of scoring the ability, and

(6) The rule of terminating the test.

The process of developing the item pool and the fitting for 3PLM were explained in the first sections of the research report. Below, we will summarise the rules 3, 4, 5, and 6 we implemented in creating our CAT algorithm.

Starting Rule: At the start of the test, an item with difficulty parameter between -0.50 and +0.50 is randomly selected from the pool and appears on the screen in front of the responder.

Ability Estimation: In estimating the ability level of responders, we preferred to use the "Maximum Likelihood-ML" method. The starting point for the ability of a responder is estimated when they answered at least 5 items, and gave at least one correct and one incorrect answer. In the application we developed, when we faced with the problem of having "more than one highest values" –a problem which is often encountered in 3PLM– it is programmed that until θ (the estimate of examinee ability) reaches an appropriate range in the ML ability estimation [-4, +4], the algorithm will continue to present a more difficult item when the responder gives a correct answer, and presents an easier item when the responder gives an incorrect answer. When the θ estimation goes out of the range between -4 and +4, the iteration of ability estimation continues receding from the θ value. To solve this problem, we used the "starting θ" formula suggested by Hambleton and Swaminathan (1985), when the first ability estimation is calculated and the estimation value of θ went out of the range of -4 and +4. The calculation of the starting ability estimation θ_{0k} is given in Equality 1 –k indicates the responder, n indicates the number of items, and r indicates the total correct answers.

$$\theta_{0k} = ln[\frac{r_a}{n - r_a}] \qquad (1)$$

Item Selection: Until the fifth answer, if the responder gives correct answer to the current item, they are presented a more difficult item, and if the responder gives incorrect answer to the current item, they are presented an easier item, selected from the pool. In the ML method, it is not possible to estimate the ability level of the responder if they give either all correct or all incorrect answers to the items. For this reason, until the responder will submit at least one incorrect or one correct answer, they will be continued to be presented items according to their difficulty level, as described for the first five items. After the first ability estimation is done, to be able to find the most informative item at that ability level, the item information function is calculated for each item. In the CAT application, we have used *Maximum Fisher Information (MFB)* method as item selection method. Here, the responder is presented the item that gives the maximum information at that ability level, selected from the items in the pool, for which the responder did not submit an answer before. In every step following this stage, the ability level of the responder is re-estimated and the cycle of delivering the item that gives the maximum information continues until the test is terminated.

After every answer submitted by the responder, their ability estimation, test information [TI(θ)], standard error [SE(θ)] are calculated and saved/updated in the database. If the test is interrupted for any reason, the responder can resume the test whenever they log in the system next time.

Termination Rule: We used two termination rules in the test. As the answers are received, the difference between the standard errors related to the ability gets smaller –indicating that the ability has reached a certain stability. When the difference between the standard errors related to the last two ability estimates is smaller than 0.01, the test is terminated –this is the first termination rule. Meanwhile, the test is also terminated when the standard error related to the ability falls below 0.30 –the second termination rule.

Elimination of Frequent Item Exposure and Content Control: Frequent item exposure means that the CAT app, if it is not limited, tends to select and deliver the statistically good items more often, a situation which causes some items to be used much more than others (Rudner, 1998). In our CAT app, we have not used any method to eliminate the frequent item exposure or to control the test content.

When *SesliTest* was completed, it has been tested on three students with visual impairment. At the end of the trial testing, we defined the problems and took opinions of the students, and improved the app so that the responders can enter their answers and replay the voiced items by using only the numerical keys. The functions, which allow students to listen to the question again, select their choice, submit their answers, change their answers and go to the next question, have been assigned to easily accessible numerical keys of the computer keyboard. The assigned keys and their functions in *SesliTest* are as follows:

Space Bar: It is used for listening the question again.

Numerical Keys: 1, 2, 3, and 4 are used to enter the choices of A, B, C, and D, respectively.

Return Key: Depending on the voiced instructions, it is used in many pages for operations such as starting the test, entering the answer, submitting the entered answer, and passing onto the next page.

Arrow Keys: They are used for fast forwarding, rewinding, pausing and resuming while listening to the items. The responder presses to the "right" key to skip to a five second later point; to the "left" key to rewind; to the "down" key to pause; and to the "up" key to resume the recording.

Item Parameters Estimated in 3PLM

The item parameters for the 166 items that fit to 3PLM have been estimated. As can be seen in Addendum 3; the item discrimination parameters (a) vary between 0.43 and 2.08; the item difficulty parameters (b) vary between -2.91 and 1.54; and the item pseudo-guessing parameters (c) vary between 0.11 and 0.38. The item discrimination parameters are set to a mean of 0.96; the mean value of the item difficulty parameters is 0.23; while the item pseudo-guessing parameters are set to a mean of 0.21. The -2 Log Likelihood value of the 166 items is obtained as 115305.9733. And we attained the characteristic curves that explain the relationship between the item parameters and the ability parameters (θ). Two examples of the obtained item characteristic curves are shown in Figure 2 and Figure 3.

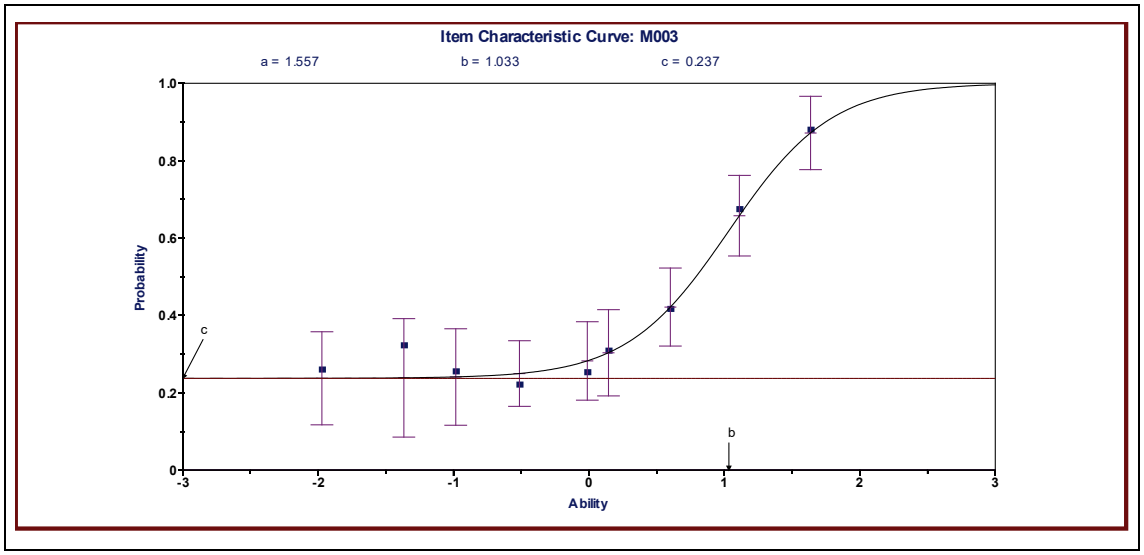


Figure 2. The Item Characteristic Curve of Item 3

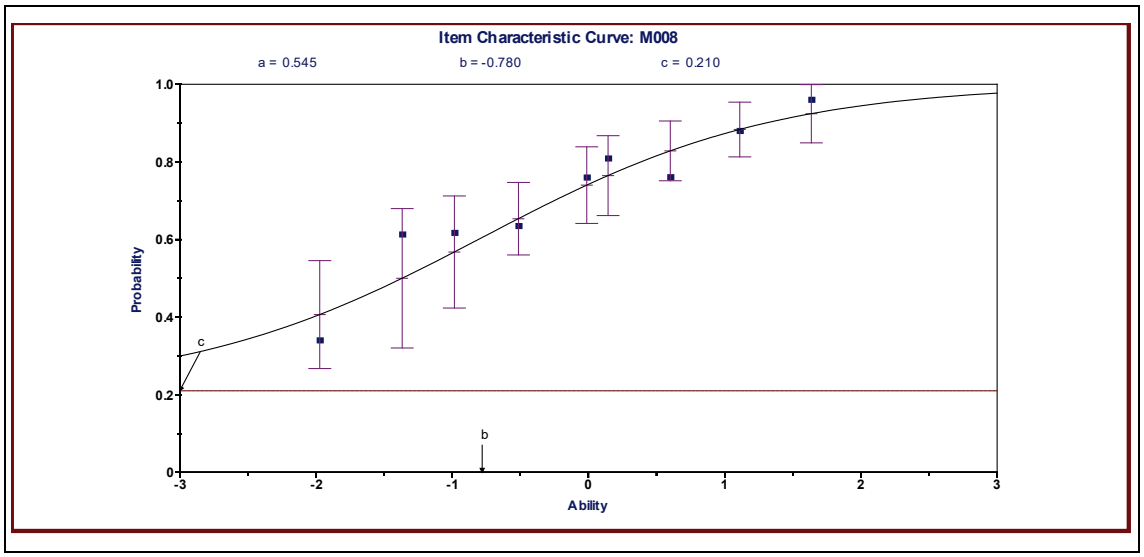


Figure 3. The Item Characteristic Curve of Item 8

When we examine the sample item characteristic curves given in Figure 2 and Figure 3, we see that the item discrimination parameter of item 3 (a_3 = 1.557) is higher than the item discrimination parameter of item 8 (a_8 = 0.545). A low *a* parameter value means that the item has a low discrimination and a high *a* parameter value means that the item has a high discrimination. This shows that the item 3 can distinguish the listening comprehension abilities of the students better than other items. When considered that the parameter a takes a value between 0.00 and 2.00 in practice, and when we evaluate the steepness of the item characteristic curve's slope, we can say that the discrimination of item 3 is very high. On the other hand, when we examine the item difficulty parameters, we see that the item difficulty parameter of item 3 (b_3 = 1.033) is bigger than the item difficulty parameter of item 8 (b_8 = -0.780). A low *b* parameter value indicates that the item is easy and a high *b* parameter value indicates that the item is difficult. This suggests that item 3 is a more difficult item than item 8. In other words, while the ability level that item 3 best measures is 1.033; the ability level best measured by item 8 is -0.780. The examination of the items' pseudo-guessing parameters gave us that the chance of answering item 3 by just guessing is higher than that of item 8. This suggests that the confusing distractors of item 8 work better.

Descriptive Statistics of the Test

As the Listening Comprehension Ability Test is a multiple choice test scored with 1-0, the lowest score achieved from the items, delivered in the app, is 0, while the highest score is 166. According to the test statistics estimated with Classical Test Theory (CTT), the arithmetic mean of the distribution of the students' scores is 93.83, their median is 92.00 and their peak value is 64.00. It is seen that the coefficient of skewness is 0.06 and the coefficient of kurtosis is -1.05. Since the coefficient of kurtosis is -1.05, we can say that the distribution is partly flat. The distribution's low kurtosis indicates that the groups display a heterogenic structure in terms of the measured property. The lowest achieved score is 30.00, while the highest is 161.00. The test range is 131.00 which is close to the maximum range of 166.00 – another sign of the heterogenic structure of the group. The mean and the median being very close to each other, while having a low peak value may appear as a deviation from the normal distribution. However, since the coefficient of skewness is within the interval of ±1.00, it can be accepted that the scores have not deviated too much from the normal distribution (Büyüköztürk, Çokluk, & Köklü, 2012) and that the test scores have got a near-normal distribution. This can also be seen in the histogram shown in Figure 4 depicting the test scores distribution.

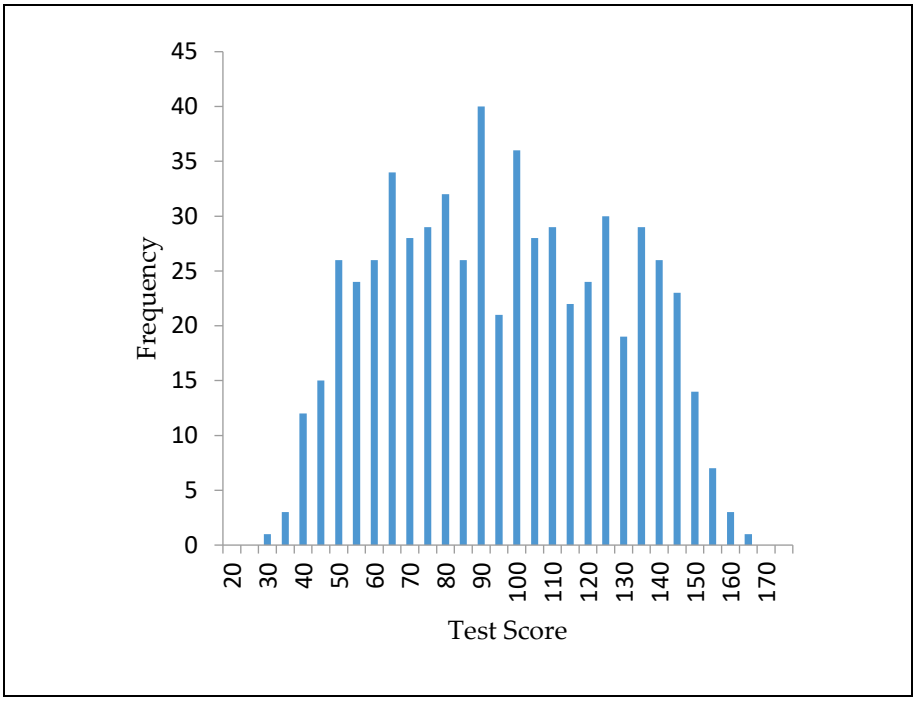


Figure 4. The Distribution Graphic of Listening Comprehension Test Scores

We calculated Kuder-Richardson 20 (KR-20) reliability coefficient, indicating the internal consistency of the Turkish Listening Comprehension Test and found 0.96. This result suggests that the items of the test are homogeneous in terms of the property they measure. We found that the marginal reliability coefficient is also 0.96. These two results can be interpreted as strong proofs that the test is reliable.

Discussion, Conclusion and Suggestions

In this study, a CAT software has been developed. All components of the software like instructions, warnings, items etc. are fully audio-based. For this aim, we created an item pool consists 166 audible items to assess the level of listening comprehension levels of students with visual impairment. Software development steps and evidences about test validity and reliability were reported in the study. Since analyzes proved test items are compatible with 3PLM, parameter estimates and software development studies were made using 3PLM. We made our CAT software multimedia design considering Mayer's (2005) multimedia guidelines. A group of students with visual impairment tried the software and test during the software and item development phase. As a result, a valid and reliable CAT has been developed for students with visual impairment at secondary school level.

The principle of providing students with special needs a life they can live independently, as emphasised in the legal documents prepared for students with special needs (MEB, 2006a), is not well observed with the practice of providing the visually impaired students human readers when they sit for tests. As also can be seen in the literature, the human reader practice, which causes the students to be "dependent on the reader", is also seen problematic by the visually impaired students themselves (Abell & Lewis, 2005; Karabay, 2016; Sánchez & Espinoza, 2012; Şenel, 2015; Tavşancıl et al., 2012). Voiced reading accommodations such as recorded human reader applications or machine readers used in computerized mediums, can free these students from being dependent on others during taking their tests (Allman, 2009). In this regard, the CAT application created as part of this study can be a significant step in meeting one of the fundamental needs of students with visual impairment and providing them the chance of "being independent individuals".

In the literature, research related to the assessment and evaluation of students with special needs is limited (Koretz & Barton, 2003). From this point of view, this research provides an important and technologic practice which has not been emphasized enough before. Additionally, findings may create awareness about the importance of validity results of the tests applied to students with visual impairment and may prove CAT is a significant tool for this purpose. SesliTest presents multi-accommodations with its specifications. We expect that this study will be viewed as an exemplary work in encouraging the use of CATs in exams administered to visually impaired students, and its assessment, software development and design stages, as explained in the report, can be a guide for future works.

As SesliTest is a short and reliable test that has positive psychological effects on responders, similar CATs can be developed for also students who have different special needs and disabilities like hearing impairment, autism, orthopaedic inability or attention deficiency. The software, developed for our study, was programmed in accordance with certain starting rules, ability estimation methods and termination rules. Seeing usages of different starting rules, ability estimation methods and termination rules, and their applications on various parametric CAT software programs can provide convenience for the researchers studying in this field.

This research focuses on design and developing of a CAT software for visually impaired students. As an extension of this research, we advise to use SesliTest or other computer-based tests that developed for similar purposes. Additionally, we may suggest using CAT especially in high-stake tests when its advantages are considered. This study is a first step for similar studies. We developed a CAT software for listening comprehension skills. Various item pools can be developed to measure different skills.

We have observed through the test and CAT software phase and in practices that getting answers by accessible keyboard keys and presenting re-listening options is very helpful for the students. Therefore, we offer to use numeric keyboard, and special keys like space and enter keys for easy access to software. We used the most frequently used keys on Q keyboard for operating and getting answers. Meanwhile, a special hardware (keyboard/mouse) specifically designed for students with visual impairment, may help to increase the functionality of the CAT software.

References

Abedi, J., Leon, S., & Kao, J. (2007). *Examining differential distract or functioning in reading assessments for students with disabilities*. Minneapolis: University of Minnesota, Partnership for Accessible Reading Assessment.

Abell, M., & Lewis, P. (2005). Universal design for learning: a statewide improvement model for academic success. *Information Technology and Disabilities Journal E-Journal*, *11*(1). Retrieved from http://itd.athenpro.org/volume11/number1/abell.html

Accountability and Curriculum Reform Effort. (2010). *Computerized adaptive testing: How CAT may be utilized in the next generation of assessments*. Retrieved from http://www.ncpublicschools.org/docs/acre/publications/2010/publications/20100716-01.pdf

Allman, C. B. (2009). *Making tests accessible for students with visual impairments: A guide for test publishers, test developers, and state assessment personnel* (4th ed.). Louisville, Kentucky: American Printing House for the Blind, Inc.

Almond, P. J., Lehr, C., Thurlow, M. L., & Quenemoen, R. (2002). Participation in large scale state assessment and accountability systems. In G. Tindal & T. M. Haladyna, (Eds.), *Large-scale assessment programs for all students: Validity, Technical Adequacy, and Implementation* (pp. 341-370). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1998). *Eğitimde ve psikolojide ölçme standartları* (S. Hovardaoğlu ve N. Sezgin, Trans.). Ankara: Türk Psikologları Derneği ve Öğrenci Seçme ve Yerleştirme Merkezi Yayını.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan Publishing Company.

Bennett, R. E. (1999). Computer-based testing for examinees with disabilities: on the road to generalized accommodation. In S. J. Messick (Ed.), *Assessment in higher education: Issues of access, quality, student development, and public policy* (pp. 181-191). Mahwah, NJ: Lawrence Erlbaum Associates

Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodations: Effects on multiple-choice reading and math items* (Technical Report 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Büyüköztürk, Ş., Çokluk, Ö., & Köklü, N. (2012). *Sosyal bilimler için istatistik* (10th ed.). Ankara: Pegem Akademi Yayınları.

Center for Evaluation, Selection and Placement. (2013). *Öğrenci Seçme ve Yerleştirme Sistemi (ÖSYS) kılavuzu*. Ankara: Ölçme Seçme ve Yerleştirme Merkezi Yayınları.

Clapper, A. T., Morse, A. B., Thompson, S. J., & Thurlow, M. L. (2005). *Access assistants for state assessments: A study of state guidelines for scribes, readers, and sign language interpreters* (Synthesis Report 58). Minneapolis, MN: National Center on Educational Outcomes, University of Minnesota.

Clark, L. (2004). Computerized adaptive testing: Effective measurement for all students. *Technological Horizons in Education Journal*, *31*(10), 14-16.

Crocker, L., & Algina, J. (1986). *Introduction classical and modern test theory.* New York: Harcourt Brace Javonovich College Publishers.

Doğan, N. (2002). *Klasik test kuramı ve örtük özellikler kuramının örneklem bağlamında karşılaştırılması* (Unpublished doctoral dissertation). Hacettepe University, Institute of Social Sciences, Ankara.

Educational Test Service. (2014). *Bulletin supplement for test takers with disabilities or health-related needs*. GRE and TOEFL Tests, The Praxis Series, Paraproand School Leadership Series Assesments.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* New Jersey: Lawrence-Earlbaum Associates.

Erdem, R. (2017). Students with special educational needs and assistive technologies: A literature review. *The Turkish Online Journal of Educational Technology*, *16*(1), 128-146.

Haladyna, T. M., & Downing, S. M. (Eds.). (2011). *Handbook of test development*. Routledge.

Hambleton, R. K. (1990). Item response theory: Introduction and bibliography. *Psicothema*, *2*(1), 97-107.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response teory: Principles and applications*. USA: Kluwer Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* California: Sage Publications.

Hausler, J., & Sommer, M. (2008). The effect of success probability on test economy and self confidence in computerized adaptive tests. *Psychology Science Quarterly*, *50*(1), 75-87.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, *26*, 44-52.

Higgins, J., & Katz, M. (2013). A comparison of audio representations of mathematics content. *Journal of Special Education Technology*, *28*(3), 59-66.

Johnstone, C. J., Altman, J., & Thurlow, M. (2006). *A state guide to the development of universally designed assessments*. Minneapolis: University of Minnesota, National Centeron Educational Outcomes.

Karabay, E. (2016). *Canlı okuyucu ve bilgisayar destekli okumanın görme engelli öğrencilerin test başarıları üzerindeki etkilerinin karşılaştırılması* (Unpublished doctoral dissertation). Ankara University, Institute of Educational Sciences, Ankara

Kelecioğlu, H. (2001). Örtük özellikler teorisindeki b ve a parametreleri ile klasik test teorisindeki p ve r istatistikleri arasındaki ilişki. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *20*, 104-110.

Kezer, F. (2013). *Bilgisayar ortamında bireye uyarlanmış test stratejilerinin karşılaştırılması* (Unpublished doctoral dissertation). Ankara University, Institute of Educational Sciences, Ankara.

Kingsbury, G. G., & Hauser, C. (2004). *Computer adaptive testing and the No Child Left Behind Act*. Paper presented at the annual meeting of the American Educational Research Association, San Diego CA. Retrieved from http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/ki04-01.pdf

Koretz, D. M., & Barton, K. (2003). *Assessing students with disabilities: issues and evidence* (CSE Technical Report. No 587). National Center for Research on Evaluation, University of California, Los Angeles.

Kutlu, Ö., & Karakaya, İ. (2004). *Orta Öğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınavının (OKÖSYS) faktör yapılarına ilişkin bir araştırma.* Paper presented at XIII. Ulusal Eğitim Bilimleri Kurultayı, Malatya.

Kutlu, Ö., Bilican, S., & Yıldırım, Ö. (2010). *İlköğretim beşinci sınıf öğrencilerinin okuduğunu ve dinlediğini anlama puanlarının farklı bilişsel düzeylere göre incelenmesi*. Paper presented at III. Uluslararası Türkçenin Eğitimi-Öğretimi Kurultayı, Dokuz Eylül University, İzmir.

Laitusis, C. C., Buzick, H., Stone, E., Hansen, E., & Hakkinen, M. (2012). *Literature review of testing accommodations and accessibility tools for students with disabilities*. Princeton, NJ: Educational Testing Service.

Laitusis, C. C., Cook, L. L., Buzick, H. M., & Stone, E. (2011). Adaptive testing options for accountability assessments. In M. Russell (Ed.), *Assessing students in the margins: Challenges, strategies and techniques* (pp. 291-310). Charlotte, NC: Information Age Publishing.

Mayer, R. E. (Ed.). (2005). *The Cambridge handbook of multimedia learning.* New York: Cambridge University Press.

Mertler, C. A. & Vannatta, R. A. (2005). *Advanced and multivariate statistical methods: practical application and interpretation*. Glendale, CA: Pyrczak Publishing.

Ministry of National Education. (2003). *Uluslararası okuma becerilerinde gelişim projesi (PIRLS) 2001 ulusal raporu*. Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı. Ankara: MEB Publishing.

Ministry of National Education. (2006a). *Özel eğitim hizmetleri yönetmeliği*. Ankara: MEB Publishing.

Ministry of National Education. (2006b). *İlköğretim Türkçe dersi (6, 7, 8. sınıflar) öğretim programı*. Ankara: MEB Publishing.

Ministry of National Education. (2012). *Ortaöğretim kurumlarına geçiş sistemi seviye belirleme sınavı başvuru kılavuzu*. Ankara: MEB Publishing.

Ministry of National Education. (2013). *Ortaöğretim kurumlarına geçiş sistemi seviye belirleme sınavı başvuru kılavuzu*. Ankara: MEB Publishing.

Minnema, J., Thurlow, M., Bielinski, J., & Scott, J. (2000). *Past and present understandings of out-of-level testing: A research synthesis* (Out-of-Level Testing Report 1). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

National Center for Learning Disabilities. (2005). No child left behind: Determining appropriate assessment accommodations for students with disabilities.

Özbaşı, D. (2014). *Bilgisayar okuryazarlığı testinin bilgisayar ortamında bireye uyarlanmış test olarak uygulanabilirliğine ilişkin bir araştırma* (Unpublished doctoral dissertation). Ankara University, Institute of Educational Sciences, Ankara.

Özbay, M. (2005). *Bir dil becerisi olarak dil eğitimi*. Ankara: Akçağ Yayınları.

Papadopoulos, K. S., & Goudiras, D. B. (2005). Accessibility assistance for visually-impaired people in digital texts. *British Journal of Visual Impairment*, *23*(2), 5-83.

Raiche, G., & Blais, J. (2002). *Practical considerations about expected a posteriori estimation in adaptive testing: adaptive a priori, adaptive correction for bias and adaptive integration interval*. Retrieved from ERIC databases (ED464110).

Rudner, L. M. (1998). *An on-line, interactive computer adaptive testing mini tutorial.* Retrieved from http://echo.edres.org:8080/scripts/cat/catdemo.htm

Russell, M., Higgins, J., & Hoffmann, T. (2009). Meeting the needs of all students: A universal design approach to computer-based testing. *Innovate: Journal of Online Education*, *5*(4).

Sánchez, J., & Espinoza, M. (2012). Chilean higher education entrance examination for learners who are blind. In Sharkey, P., & Klinger, E. (Eds.), *Proceedings of the 9th international conference on disability, virtual reality and associated technologies* (pp. 409-418), Laval, France.

Şenel, S. (2015). Görme engelli öğrencilerin üniversite giriş sınavı deneyimleri. *Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü Eğitim Araştırmaları Dergisi*, *1*(1), 1-17.

Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodation on test performance: A review of the literature.* Center for Educational Assessment Research Report no: 485. Amherst, MA: School of Education, University of Massachusetts Amherst.

Stone, E., & Davey, T. (2011). *Computer-adaptive testing for students with disabilities: A review of the literature. Research Report*. Educational Testing Service, Princeton, New Jersey.

Tavşancıl, E., Uluman, M., & Furat, E. (2012). *Görme engelli öğrencilerin üniversite giriş sınavında karşılaştığı sorunlar ve çözüm önerileri*. Paper presented at III. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi, Bolu.

Thompson, N. A. (2010). *Adaptive testing: is it right for me?* Minesota: University of Cincinnati. Assessment Systems Corporation. Retrieved from http://www.assess.com/docs/Thompson_(2010)_-_Adaptive_Testing_Right.pdf

Thurlow, M., Lazarus, S. S., Albus, D., & Hodgson, J. (2010). *Computer-based testing: Practices and considerations* (Synthesis Report No. 78). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Tian, J., Miao, D., Zhu, X., & Gong, J. (2007). An introduction to the computerized adaptive testing. *Us-China Education Review*, *4*(1), 26.

Tindal, G. (1998). *Models for understanding task comparability in accommodated testing*. Eugene, OG: Behavioral Research and Teaching.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 361-375.

Yaman, F., Dönmez, O., Avcı, E., & Kabakçı Yurdakul, I. (2016). İşitme Engelli Öğrencilerin Okuma-Yazma Eğitiminde Mobil Uygulama Kullanımı. *Eğitim ve Bilim*, *41*(188), 153-174. doi:10.15390/EB.2016.6687

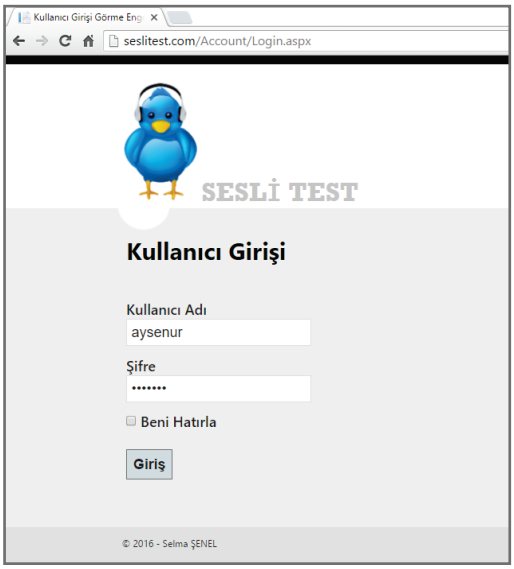
Yıldırım, H., Çömlekoğlu, G., & Berberoğlu G. (2003). Milli Eğitim Bakanlığı özel okullar sınavı verilerinin madde tepki kuramı modellerine uyumu. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *24*, 159-168.

Appendix 1. Table of Test's Scope – Expression Unit – Intellectual Levels

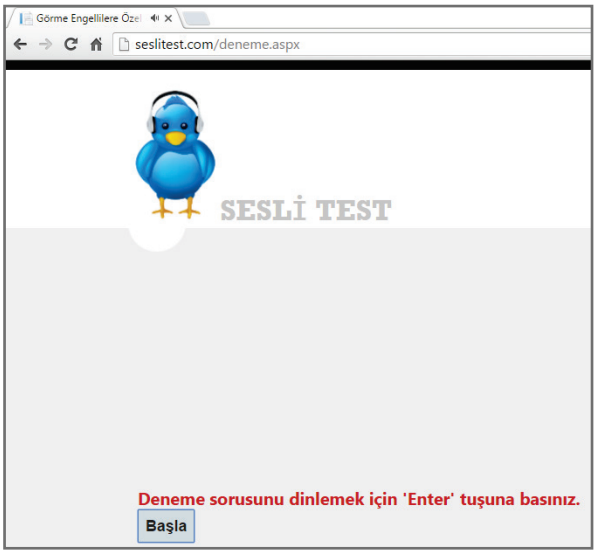
Sub-Field	Acquisitions	Expression Unit	Intellectual Level	Number of Items
A. Understanding and analysing the listened topics	A1. Dinlenenin bağlamından hareketle kelime ve kelime gruplarının anlamlarını çıkarır.	Word	1.level	8
	A2. Dinlediklerindeki anahtar kelimeleri fark eder.	Word	1.level	8
	A3. Dinlediklerinin konusunu belirler.	Paragraph	2.level	8
	A4. Dinlediklerinin ana fikrini/ana duygusunu belirler.	Paragraph	1.level	15
	A5. Dinlediklerindeki yardımcı fikirleri/duyguları belirler.	Paragraph	2.level	20
	A6. Dinlediklerindeki olay, yer, zaman, şahıs, varlık kadrosu ve bunlarla ilgili unsurları belirler.	Paragraph	1.level	9
	A7. Dinlediklerinde sebep-sonuç ilişkilerini belirler.	Paragraph	2.level	7
	A8. Dinlediklerinde amaç-sonuç ilişkilerini belirler.	Paragraph	2.level	8
	A9. Dinlediklerindeki örtülü anlamları bulur.	Sentence	2.level	11
	A10. Dinlediklerini kronolojik sıra ve mantık akışı içinde özetler.	Paragraph	2.level	7
	A11. Dinlediklerine ilişkin sorular oluşturur.	Paragraph	2.level	8
	A12. Dinlediklerine ilişkin sorulara cevap verir.	Paragraph	2.level	9
	A13. Dinlediklerinde yer alan öznel ve nesnel yargıları ayırt eder.	Sentence	3.level	6
	A14. Dinlediklerine ilişkin karşılaştırmalar yapar.	Paragraph	3.level	8
	A15. Kendisini şahıs ve varlık kadrosunun yerine koyarak olayları, duygu, düşünce ve hayalleri yorumlar.	Paragraph	3.level	7
	A16. Dinlediklerinde ortaya konan sorunlara farklı çözümler üretir.	Paragraph	3.level	5
	A17. İpuçlarından hareketle dinlediklerine yönelik tahminlerde bulunur.	Paragraph	3.level	9
	A18. Dinlediklerinin öncesi ve/veya sonrasına ait kurgular yapar.	Paragraph	4.level	7
	A19. Dinlediklerinin başlığı/adı ile içeriği arasındaki ilişkiyi ortaya koyar.	Paragraph	3.level	7
	A20. Dinlediği metne farklı başlıklar bulur.	Paragraph	3.level	5
	A21. Görsel/işitsel unsurlarla dinledikleri arasında ilgi kurar.	Paragraph	3.level	5
	A22. Şiir dilinin farklılığını ayırt eder.	Paragraph	3.level	5
	A23. Şiirin kendisinde uyandırdığı duyguları ifade eder.	Paragraph	3.level	6
B. Evaluating the listened topics	B1. Dinlediklerini dil ve anlatım yönünden değerlendirir.	Paragraph	4.level	8
	B2. Dinlediklerini içerik yönünden değerlendirir.	Paragraph	4.level	9
C Enriching the lexicon	C1. Kelimeler arasındaki anlam ilişkilerini kavrayarak birbiriyle anlamca ilişkili kelimelere örnek verir.	Word	2.level	9
	C2. Aynı kavram alanına giren kelimeleri, anlam farklılıklarını dikkate alarak kullanır.	Word	3.level	7
	C3. Dinlediklerinde geçen kelime, deyim ve atasözlerini cümle içinde kullanır.	Sentence	3.level	11

Appendix 2. Screen Shots from Audio-Based CAT Application (seslitest.com)

Home Page of Seslitest



User Login Page



The Sample Item Screen

Appendix 3. Parameters of 166 Items Estimated in 3PLM

No	a	SE_a	b	SE_b	c	SE_c	No	a	SE_a	b	SE_b	c	SE_c
1	0,50	0,09	-2,91	0,50	0,20	0,09	44	0,91	0,24	1,32	0,16	0,28	0,04
2	0,78	0,19	1,20	0,18	0,25	0,05	45	0,91	0,14	0,17	0,16	0,23	0,06
3	1,59	0,31	1,02	0,08	0,24	0,03	46	0,70	0,09	-1,07	0,21	0,17	0,07
4	0,47	0,07	-0,86	0,30	0,17	0,08	47	1,26	0,18	-0,54	0,14	0,26	0,07
5	0,65	0,10	-1,80	0,29	0,20	0,09	48	1,24	0,18	0,03	0,12	0,28	0,05
6	0,70	0,11	-0,47	0,24	0,23	0,08	49	0,75	0,14	-0,03	0,23	0,30	0,08
7	0,55	0,09	-1,63	0,33	0,21	0,09	50	1,08	0,21	0,99	0,12	0,26	0,04
8	0,56	0,09	-0,75	0,29	0,21	0,09	51	1,08	0,18	-0,65	0,16	0,22	0,08
9	0,54	0,08	-0,56	0,27	0,18	0,08	52	0,69	0,12	0,06	0,22	0,22	0,07
10	0,76	0,14	0,88	0,16	0,19	0,05	53	1,27	0,23	1,11	0,10	0,23	0,03
11	0,70	0,17	1,43	0,19	0,20	0,04	54	1,27	0,18	0,03	0,11	0,22	0,05
12	0,68	0,10	-0,22	0,21	0,19	0,07	55	1,02	0,15	-0,86	0,18	0,24	0,08
13	0,63	0,11	0,41	0,19	0,16	0,06	56	0,79	0,09	-0,39	0,14	0,11	0,05
15	0,66	0,12	0,41	0,20	0,20	0,06	57	1,08	0,15	-0,44	0,13	0,18	0,06
16	0,59	0,08	-0,73	0,27	0,21	0,08	58	0,76	0,14	0,53	0,17	0,22	0,06
17	0,65	0,12	0,24	0,25	0,27	0,07	59	0,79	0,11	0,37	0,13	0,12	0,05
18	0,48	0,09	0,81	0,24	0,15	0,06	60	0,68	0,11	0,03	0,23	0,23	0,07
19	0,54	0,08	-0,88	0,30	0,20	0,09	61	0,83	0,14	0,28	0,17	0,22	0,06
20	0,69	0,10	-0,51	0,23	0,21	0,08	62	0,76	0,13	0,38	0,17	0,20	0,06
21	0,77	0,10	-0,76	0,19	0,17	0,07	63	1,21	0,17	-0,13	0,11	0,21	0,05
22	0,99	0,24	1,14	0,14	0,26	0,04	64	0,98	0,14	-0,52	0,17	0,22	0,07
23	0,98	0,14	-0,88	0,19	0,25	0,08	65	0,46	0,08	0,38	0,28	0,17	0,07
24	0,58	0,12	0,62	0,25	0,23	0,07	66	0,53	0,08	-0,07	0,22	0,14	0,06
25	0,84	0,17	0,98	0,13	0,17	0,04	67	1,04	0,13	-0,28	0,13	0,16	0,06
26	0,82	0,14	-0,41	0,22	0,27	0,08	68	0,97	0,15	0,17	0,13	0,20	0,05
27	0,99	0,12	-0,61	0,13	0,14	0,06	69	0,85	0,14	-0,48	0,22	0,27	0,08
28	0,72	0,12	0,07	0,19	0,19	0,07	70	0,98	0,16	0,35	0,14	0,25	0,05
29	1,06	0,14	-1,05	0,18	0,22	0,08	71	1,55	0,23	-0,42	0,11	0,27	0,06
30	0,70	0,08	-0,87	0,19	0,14	0,06	72	0,76	0,09	-0,33	0,17	0,15	0,06
31	0,99	0,17	0,50	0,14	0,25	0,05	73	0,90	0,15	0,32	0,15	0,21	0,06
32	0,95	0,12	-0,32	0,13	0,14	0,05	74	0,86	0,10	-0,38	0,14	0,13	0,05
33	1,02	0,14	-0,59	0,15	0,19	0,07	75	0,77	0,10	-0,21	0,16	0,14	0,06
34	0,93	0,10	-0,41	0,13	0,13	0,05	77	1,14	0,18	-0,47	0,16	0,28	0,07
35	0,75	0,14	0,51	0,17	0,21	0,06	78	0,81	0,15	0,59	0,16	0,21	0,05
36	0,90	0,11	-0,16	0,12	0,11	0,05	79	1,28	0,24	1,00	0,10	0,26	0,03
38	0,69	0,12	0,80	0,16	0,14	0,05	80	1,58	0,32	1,20	0,09	0,26	0,03
39	0,92	0,16	0,38	0,14	0,23	0,05	81	0,93	0,17	0,12	0,17	0,29	0,06
40	0,73	0,13	-0,10	0,23	0,26	0,08	82	1,00	0,14	0,24	0,12	0,16	0,05
41	0,56	0,11	0,96	0,21	0,16	0,06	83	0,85	0,11	-0,30	0,16	0,18	0,06
42	0,67	0,10	-0,15	0,23	0,22	0,07	84	1,12	0,17	0,26	0,12	0,26	0,05
43	1,13	0,19	0,63	0,11	0,21	0,04	85	0,55	0,10	0,62	0,23	0,18	0,06

No	a	SE_a	b	SE_b	c	SE_c	No	a	SE_a	b	SE_b	c	SE_c
86	1,52	0,24	-0,27	0,11	0,32	0,05	131	0,57	0,08	-0,75	0,24	0,16	0,07
87	0,54	0,11	0,67	0,26	0,22	0,07	132	1,32	0,26	0,94	0,10	0,25	0,03
88	0,92	0,15	0,32	0,15	0,24	0,05	133	1,19	0,26	0,80	0,13	0,38	0,04
89	0,88	0,17	0,65	0,15	0,27	0,05	134	1,02	0,16	0,10	0,14	0,25	0,06
90	0,66	0,12	0,48	0,20	0,20	0,06	135	0,89	0,12	-0,12	0,14	0,15	0,06
91	1,15	0,14	-0,80	0,14	0,19	0,07	136	0,93	0,13	-0,09	0,15	0,21	0,06
92	0,89	0,13	0,38	0,12	0,14	0,05	137	0,65	0,12	0,70	0,19	0,18	0,06
93	0,58	0,09	-0,53	0,26	0,20	0,08	138	1,69	0,29	0,54	0,08	0,25	0,03
94	0,97	0,13	-0,62	0,17	0,22	0,07	139	1,19	0,21	0,49	0,11	0,27	0,04
95	0,43	0,08	0,47	0,31	0,18	0,07	140	0,51	0,11	0,51	0,30	0,23	0,08
96	0,99	0,20	1,04	0,12	0,21	0,04	141	1,63	0,28	0,80	0,08	0,23	0,03
97	1,16	0,24	1,07	0,11	0,23	0,03	142	1,21	0,15	-0,23	0,10	0,13	0,05
98	1,03	0,13	-0,50	0,15	0,20	0,07	143	1,00	0,17	0,47	0,14	0,27	0,05
100	0,97	0,16	0,79	0,11	0,16	0,04	144	0,87	0,14	0,59	0,13	0,18	0,05
101	1,25	0,20	0,57	0,09	0,19	0,04	145	1,96	0,43	1,08	0,07	0,21	0,02
102	0,91	0,14	-0,31	0,18	0,23	0,07	146	1,14	0,18	0,30	0,11	0,22	0,05
103	0,62	0,10	0,46	0,20	0,16	0,06	147	1,17	0,28	1,54	0,14	0,18	0,03
104	1,22	0,26	1,14	0,10	0,22	0,03	148	2,08	0,48	1,06	0,07	0,27	0,03
105	1,00	0,14	-0,21	0,13	0,17	0,05	149	1,85	0,30	1,03	0,07	0,22	0,02
107	0,74	0,15	0,67	0,19	0,26	0,06	150	1,59	0,40	1,21	0,10	0,32	0,03
108	0,58	0,08	-0,68	0,25	0,17	0,08	151	1,13	0,16	-0,03	0,13	0,26	0,05
109	0,95	0,12	-0,06	0,12	0,13	0,05	152	0,89	0,13	0,17	0,13	0,15	0,05
110	0,58	0,08	0,02	0,20	0,14	0,06	153	1,12	0,18	0,70	0,10	0,19	0,04
111	0,90	0,11	-0,01	0,12	0,13	0,05	154	0,60	0,11	0,52	0,22	0,20	0,07
112	0,86	0,13	0,54	0,13	0,17	0,05	155	0,82	0,18	1,06	0,15	0,22	0,05
113	1,11	0,17	0,14	0,12	0,20	0,05	156	1,14	0,21	0,06	0,14	0,31	0,06
114	1,36	0,23	-0,36	0,11	0,21	0,06	157	1,01	0,20	0,77	0,12	0,25	0,04
115	1,25	0,17	-0,39	0,12	0,21	0,06	158	1,00	0,19	0,52	0,14	0,29	0,05
116	0,92	0,17	0,62	0,15	0,27	0,05	159	1,03	0,12	-0,20	0,11	0,14	0,05
117	0,98	0,10	-0,23	0,11	0,12	0,04	160	0,83	0,15	0,99	0,13	0,14	0,04
118	1,87	0,29	0,27	0,07	0,26	0,04	161	1,14	0,22	0,45	0,14	0,35	0,05
119	1,18	0,18	0,69	0,10	0,21	0,04	162	0,73	0,18	1,21	0,19	0,25	0,05
120	1,33	0,22	0,72	0,09	0,20	0,03	163	1,48	0,23	0,42	0,09	0,25	0,04
121	0,83	0,16	0,58	0,16	0,22	0,05	164	1,01	0,19	0,63	0,14	0,28	0,05
123	0,80	0,13	0,22	0,19	0,26	0,06	165	1,17	0,22	0,82	0,11	0,23	0,04
124	1,00	0,23	1,04	0,15	0,34	0,04	166	1,27	0,34	1,46	0,13	0,24	0,03
125	1,22	0,20	0,28	0,12	0,27	0,05	167	0,69	0,17	1,22	0,20	0,27	0,05
126	1,31	0,26	0,98	0,11	0,31	0,03	168	1,04	0,15	0,64	0,10	0,15	0,04
127	0,86	0,20	1,24	0,15	0,22	0,04	169	1,28	0,21	0,78	0,09	0,17	0,03
128	0,50	0,11	1,35	0,25	0,16	0,05	171	0,67	0,14	1,18	0,18	0,18	0,05
129	0,72	0,13	0,37	0,19	0,22	0,06	172	1,56	0,36	1,26	0,09	0,22	0,03
130	1,04	0,16	0,55	0,12	0,22	0,04							