Turkish Expressive and Receptive Language Test: I. Standardization, Reliability and Validity Study of the Receptive Vocabulary Sub-Scale

Sibel KAZAK BERUMENT[1], Ayşe Gül GÜVEN[2]

SUMMARY

Objective: A reliable, valid and original test to assess the receptive vocabulary skills of children in Turkey was not available. Thus, the purpose of the current study was to develop a receptive vocabulary test for Turkish children based on the Turkish language.

Materials and Methods: For the Receptive Vocabulary Sub-Scale (TIFALDI-RT) 242 concrete and abstract words were chosen from word frequency lists and a comprehensive Turkish Dictionary. Pilot data were collected from 648 children aged 2 to 13 from Ankara, and norm data were collected from a nationally representative sample of 3755 children.

Results: Item analysis (item difficulty, discrimination and distractor) was carried out on the pilot data and based on the results, the total item number was reduced to 157. Further, three parameter item analyses (IRT) were carried out on the norm data by using BILOG-MG (SSI, 2002), and the results indicated that the TIFALDI Receptive Vocabulary Sub-Scale could be reduced to 104 items to assess 2 to 12 year-old children's receptive vocabulary. Test-retest and internal consistency reliabilities were calculated for the whole sample and age groups separately, and all the coefficients were high. For the validity, the relationship between the WISC-R and Ankara Developmental Screening Inventory (AGTE) and Receptive Vocabulary Sub-Scale were investigated. Once again, the TIFALDI Receptive Vocabulary Sub-Scale scores were found to be significantly related to WISC-R and AGTE scores.

Conclusion: The TIFALDI Receptive Vocabulary Sub-Scale was developed on the basis of the Turkish Language and norm data were collected from a nationally representative sample. The TIFALDI-RT also had a high reliability and validity. Thus, the TIFALDI-RT can be used to assess 2 to 12 year-old children's receptive vocabulary skills.

Key words: language tests, language development, receptive language

INTRODUCTION

Language acquisition is one of the important developmental areas of a child's development. Starting from birth up to adult level language comprehension and production, children need to acquire different aspects of language like *phonology*, *syntax*, *morphology*, *semantics and pragmatics*. In language acquisition, an infants' first task is to recognize and differentiate the sound characteristics of the language they are exposed to (Levey & Polirstok 2011). In fact, at birth infants are not only able to make sound differentiations –this is referred to as categorical speech perception- of the mother's tongue but also they are able to make these differentiations in many other languages they have never heard (Werker & Tees 1984; Werker &Desjardins 1995). From birth onward, infants' categorical speech perception skills become more advanced for the language they hear, but this ability shows a significant decrease from 6 to 12 months for other languages (Polka & Werker 1994). Infants show preference to the language that their mothers speak during pregnancy (Moon et. al. 1993) and from 6 months onward they begin to comprehend the language they hear (Jusczyk &Aslin 1995). At 2 months of age, infants start cooing, at which time they produce vowel sounds and consonants, and around the age of 6 months they start babbling (Levey &

Received: 12.07.2012 - **Accepted:** 17.09.2012

[1]Middle East Technical University Department of Psychology., [2]Eskişehir Osmangazi University Department of Audiology.

E-mail: *sibel@metu.edu.tr*

Polirstok 2011). When infants become 10 months old their babbling resembles more of the language they have been exposed to, and these sound productions turn into first words (Boysson-Bardies &Vihman 1991). Usually infants produce their first meaningful words around the age of 1, but until 18 months of age, the rate of word acquisition is usually slow (Bee 2000; Fenson et. al. 1994). As infants' vocabulary size reaches aproximalety 50 words at 18 to 22 months of age, their word acquisition rates begin to get faster (Goldfield &Reznick 1990). Children can use nearly 600 words by 30 months and over ten thousand words around the age of 5 and 6 years old (Bates et. al. 1994; Herschensohn 2007).

In language acquisition, vocabulary size is one of the main indicators of semantic development (Levey &Polirstok 2011). Genetic or biologically based developmental problems, hearing loss and lack of environmental input can lead to delay or disorders of language. Problems in the acquisition of words could be a sign of problems in language development (Okalidou et. al. 2011). Therefore, vocabulary tests have an important function in the assessment of language development. Usually up to the age of two years old, assessments of children's vocabulary size is made through parental reports. The MacArthur Communicative Development Inventory (Fenson et. al. 1993) can be given as a good example to these types of tests. This scale has been translated and adapted into a number of languages like German, Japanese, French, British English (Okalidou et. al. 2011), and recently a Turkish adaptation has also been done (TİGE) (Aksu-Koç et. al. 2011).

After two years of age, the assessment of a child's language development can be done through individually administered tests. The British Picture Vocabulary Scale –BPVSIII (Dunn et. al. 2011), Peabody Picture Vocabulary Test -PPVT-IV (Dunn & Dunn 2007), Test of Early Language Development -TELD-3 (Hresko et.al. 1999), Receptive One-Word Picture Vocabulary Test -ROWPVT-4 (Martin & Brownell 2011), Preschool Language Assessment Instrument -PLAI-2 (Blank et. al. 2003), Preschool Language Scale-4 (Zimmerman et.al. 2005), Test of Word Knowledge -TOWK (Wiig & Secord 1992), Test of Reception of Grammar -TROG-2 (Bishop 2003), and the Test of Language Development -TOLD-4 (Hammill & Newcomer 2005) can be given as examples of these tests.

Due to the limited availability of language tests in Turkey, at clinical settings, sub-scales of some developmental or intelligence tests are widely used to assess language development. The Vineland Adaptive Behaviour Scale [Davranış Uyum Ölçeği] (Alpas &Akçakın 2003), Ankara Developmental Screening Inventory (Savaşır et. al. 1998), Wechsler Intelligence Scale for Children (Wechsler D, 1992)) and Denver Development Screening Scale (Anlar&Yalaz 1996) can be given as examples to these tests.

There has not been an original Turkish receptive vocabulary test developed based on the Turkish language. The Peabody Picture Vocabulary Test was adapted to Turkish in 1972 and since then has been used to assess 2 to 11 year old children's receptive vocabulary. However, there are serious concerns about the use of this test. The most important drawback is the outdated norms. For instance, when a typically developing four year old child's raw score is converted to their age equivalent, it can come up as 7 years 6 months, and standart score as 187,5. Because of this, the Turkish Peabody Picture Vocabulary Test (1972) cannot be used to assess children's receptive vocabulary. Similarly, researchers who used this test in their studies have had problems publishing their research. Therefore, there is an important need for a measurement to assess Turkish speaking children's receptive vocabulary skills at clinical settings and for research purposes.

In order to fill this gap, in 1998 development of a langauge test which was based on the Turkish language was started. Knowing that the translation of language tests has a number of shortcomings (like item difficulty differences) (Zumbo 2003) due to the nature of the languages, the goal was to develop an original test rather than adapting an existing test from another language. Accordingly, the aim of this study was to develop an original, reliable and valid Turkish Receptive Vocabulary Test with nationally representative norms for children aged 2 to 15 years.

Development of Receptive Vocabulary Test: Pilot Study

METHOD

Selecting Test Items

To form the test item pool, three different sources were used. First, words were chosen from a concrete Turkish words frequency list (Er, 1996). Second, to form the abstract and action Turkish words frequency lists, 1000 university students were given two-sided sheets with 28 of the 29 letters in the Turkish alphabet written on them. For each letter on one side of the page, participants were asked to write the first "action word" that came to their minds. For the other side of the page they were asked to write the first "abstract word" that came to their minds. From this data, the abstract and concrete word lists were made. Lastly, Püsküllüoğlu (1997) Turkish Dictionary was examined to identify words that could be represented through drawings.

From these word lists, 242 words with varying usage frequency were selected. These words were from superordinate categories of animals, clothing, food, fruit and vegetables, household items, furniture, occupations, body parts, math terms, toys, vehicles, building-house-garden sections, emotions, actions, stationary goods and utensils.

For each target word 3 alternatives from the same semantic category were selected. For instance, if a target word was an *animal,* alternatives were also *animals*, and if a target was an *emotion* word, than the alternatives were chosen from other *emotion* words. For all 242 items, the target words' placement in the card was randomly designated, and pictures of the target and alternative words were drawn by a professional artist.

Participants

Participants of the pilot study were recruited for all SES and maternal education levels from the towns and villages of Ankara city. A total of 648 children between the ages of 2 to 13 years were tested.

Procedure

For data collection the schools approval was required from the Ministry of Education. Pilot data were collected by a trained psychology graduate and final year undergraduate students. Although at the start our target age range was from 2 years to 15 years of age, a ceiling effect was observed for 13 year olds during pilot data collection, and thus the test was not given to 14 and 15 year olds. Since there were 242 items for the pilot phase and it was necessary to get data for each card, we started with children aged 6 years and above. The item difficulty analysis was carried out for the data that were gathered for 6 to 13 year olds. Based on the item difficulty results of the 6 year olds from the 242 items, 100 easy items were selected and given to 4 and 5 year olds. Finally, based on the item difficulty results of the 4 year olds, 57 easy items were selected and given to the 2 and 3 year olds.

RESULTS

Item Analysis

Item difficulty, item discrimination and distractor analyses were carried out and based on the results, some items were eliminated. Due to distractor problems - if predominantly one or two of the distractors were chosen- 52 items (friendship, throw, knife, kettle, pull, valuable, patient, chase, faucet, argue, carry, wheel, fly, fireplace, tweezers, patch, candleholder, hide, carpenter, bead, entertainment, tooth, cook, spin, coffee pot, wet, knee, tomato, book, pain, bread, painter, horseshoe, carpet, pilot, old, bite, untidy, moon, jumper, pergola, heroism, towel, obstacle, tower, shirt, screw, get on, choose, buttonhole, peak and sieve) were eliminated, due to item discrimination problems 9 items were eliminated (slid, drainer, purse, barrel, get off, scoop, leek, sparse and tie) and due to drawing imperfections –when redrawing of the picture could not solve the problem- 22 items were eliminated (axe, bunch, fridge, drawer, groom, camel, beans, drink, armchair, necklace, fear, shovel, leader, minibus, anger, clown, money, aubergine, sweet, bottle, lazy and sky). According to the pilot study's results, the Receptive Vocabulary Sub-Scale was revised by reducing the number of the items from 242 to 159 and reordering them from easiest to most difficult.

Development of Receptive Vocabulary Test: Norm Study

METHOD

Participants

To determine a nationally representative sample of 2 to 13 year olds, an application was made to the Turkish Statistical Institute (TSI) asking for a minimum of 300 children in each age group. TSI identified 2880 main and 5760 secondary addresses from the villages and towns of 61 cities of Turkey. Norm data were collected from 3755 children residing in 158 settlements between June 2007 and November 2008. Table 1 shows the number of targeted and reached houses, and the number of children tested in each city.

Data were screened for missing values or possible tester faults, and these cases were eliminated. Analyses were started with data from 3650 children (1760 female, 1819 male and 71 with no gender information) collected from 2626 houses in 61 cities.

MATERIAL

Based on the pilot study results, the Turkish Receptive Vocabulary Sub-Scale included two trials and 157 test items to be used for the evaluation of 2 to 13 year-old children's receptive vocabulary skills. Each card includes 4 pictures one of which is the target and the rest are the alternatives, all of which were drawn by a professional artist. During the administration of the trials, the children are instructed to find a picture of the target word that is told by the tester.

Procedure

Before the data collection, written permissions were taken from each city's governorship. In order to increase the participation rate, 2 to 3 weeks prior to home visits, all (8640) main and secondary home addresses identified by the Turkish Statistical Institute received a letter explaining the study and informing them about the home visit and a copy of the permission letter from the Governship's Office.

Field workers were mostly final year psychology students chosen from Middle East Technical University and from Hacettepe University, some graduate students were chosen as well. Students who wished to work on the field for data collection went through a training given by the project holders. From the successful candidates, 34 people were chosen and they formed 17 pairs. The major part of data collection was

completed by these groups, but in Istanbul the number of targeted houses was very high. Thus, field training was repeated in Istanbul for willing psychology students from local universities -Bilgi University, Koç University and Boğaziçi University- and from 12 successful candidates 6 pairs were formed.

For each city, according to the number of targeted addresses, field worker pairs were identified. Each group was given a set of materials including a test battery, demographic information form, list of addresses, parent consent forms, permission letters from the local Governship's Office as well as pencils and stickers as rewards for children.

Data were collected from targeted addresses, but if parents did not give permission, or on consequent trials if the family could not be reached, or if the given address was for a business office, secondary addresses were used. Data were collected only from children whose parents gave consent for their participation.

Analysis

For TİFALDİ Receptive Vocabulary Sub-Scale item analyses, the BILOG-MG (SSI 2002) program was used to carry out 3 parameters (item difficulty, discrimination and guess) of the Item Response Theory (IRT) analyses. To determine stopping rule, probability calculations were accomplished (Newbold et. al. 2003) and total raw score normalizations were carried out by converting raw scores to percentile ranks for each age range separately.

RESULTS

Item Analysis

For 157 items of the Receptive Language Vocabulary Sub-Scale, 3 parameters -item difficulty, discrimination and guess- for the Item Response Theory (IRT) analysis was done by

Table 1. The number of targeted and accessed houses and number of children tested in each city listed by TurkStat

City	Houses		Number of tested children	City	Houses		Number of tested children
	Target	Accessed			Target	Accessed	
Adana	90	90	135	Erzurum	15	15	21
Afyon	45	42	54	Eskişehir	30	31	41
Ankara	210	181	269	Gaziantep	45	45	88
Antalya	60	58	74	Giresun	15	15	23
Ardahan	15	15	20	Gümüshane	15	15	24
Artvin	15	15	21	Hatay	45	45	74
Aydın	45	38	48	Iğdır	15	15	27
Balıkesir	75	64	84	Isparta	30	30	40
Batman	15	12	20	İstanbul	540	440	572
Bilecik	15	12	19	İzmir	210	178	214
Bingol	15	12	14	Kahramanmaraş	45	45	70
Bitlis	15	17	34	Karabük	30	24	35
Bolu	15	12	17	Karaman	15	14	15
Burdur	15	15	15	Kastamonu	30	26	35
Bursa	120	119	161	Kayseri	45	45	58
Çorum	30	28	39	Kırıkkale	30	27	37
Denizli	60	38	55	Kırklareli	30	31	36
Diyarbakır	30	30	38	Kilis	15	15	26
Elazığ	15	15	24	Kocaeli	45	35	43
Erzincan	15	15	19	Konya	90	89	120
Kütahya	30	31	34	Samsun	45	45	66
Malatya	45	44	62	Sivas	30	30	42
Manisa	45	41	56	Tekirdağ	45	29	43
Mardin	15	11	16	Tokat	15	15	19
Mersin	60	58	89	Trabzon	45	46	84
Muğla	30	30	47	Urfa	30	30	50
Nevşehir	30	30	41	Uşak	30	28	40
Niğde	15	15	22	Van	15	15	20
Ordu	15	15	26	Yalova	30	30	49
Osmaniye	15	15	24	Yozgat	30	30	31
Sakarya	30	30	30				
TOTAL					2895	2626	3650

using BILOG-MG (SSI 2002). In order to see the item functions for particular ages, analysis was first carried out for the whole sample, then by dividing the data into two groups of 2 to 6 years and 7 to 13 years; then for 2-3-4 years, 5-6-7 years and 8 to 13 years; and lastly for each age group separately.

Based on the item difficulty results of the whole sample, items were reordered from easiest to most difficult. The mean score for **item difficulty** was — 0.47, and ranged from — 2.33 to 2.75. The mean score for **item discrimination was** 1.72, and ranged from 0.48 to 3.16. Since the sample size should be a minimum of 1000 for IRT analysis, only the grouped data analysis results were considered. When evaluating the results, age starting points were taken as a base. For instance, for 2-3-4 year olds' combined analyses, results were judged up to the 5 year olds starting point for the test administration.

Starting points for the test were decided according to the percentages of correct responses for each age. Then, item difficulty, item discrimination, and item functions graphs of each item were scrutinized to see whether there were any items to be eliminated from the test. First, items were ordered according to item difficulty calculations for 2-3-4 year olds, and items with equal difficulty rates (ring, tray, balcony, needle, cupboard, quilt, packet, stick, lean on, cut, pin, fire and push) and items with discrimination scores lower than .75 (balloon, doll, horse and skirt) were emitted. Next, from 5 year olds starting point to the test (item 33) items were reordered according to the item difficulty calculations of 5-6-7 year olds, and items with equal difficulty rates (to sweep, squirrel, coat and to press) and items with discrimination scores lower than .75 (bee, key ring, raise, help, scythe, wipe, tulip, mirror and pick up) were extracted from the test. Then, from 8 year olds starting point to the test (item 69) items were reordered according to the item difficulty calculations of 8-11 and 8-13 year olds. Inspection of the item difficulty scores indicated that items for 13 year olds were too easy. Therefore, it was decided that this Receptive Vocabulary Sub-Scale in itself, is not suitable for youngsters older than 12 years of age. Finally, IRT analyses were repeated for 8-to 12 year olds together and for items with discrimination scores lower than .65 (tea urn, water pipe, vest, freedom, lantern, tired, monument, barrel, broken, hook, trust, saw, stretch oneself, excavation, law, hook, valley, spin yarn, transfer, serenity, log, to wait and resistance). When IRT analyses were done on the combined age groups of 8 to 12 years, some of the 11 and 12 year old level items (e.g. lamp, plow, scoop, viaduct, tape-measure) had low discrimination scores. When analyses were repeated for the 11 and 12 year old level they had acceptable levels of discrimination power, thus they were kept in the test.

In sum, based on the Receptive Vocabulary Sub-Scale norm data IRT analyses, 53 items were eliminated from the test and some items' orders were changed. It has been suggested that with these revisions the Receptive Vocabulary Sub-Scale with 104 items is suitable for children aged 2 to 12. The age graded calculations for the mean **item difficulty** score of 104 items was — 0.44, ranged from –1.76 to 2.31, and the mean score for **item discrimination was** 1.08, ranged from 0.32 to 3.06. IRT analyses of 104 items for the whole sample mean score for **item difficulty** was — 0.48, ranged from — 2.31 to 2.73, while the mean score for **item discrimination was** 1.81, ranged from 0.49 to 3.12. Mean **item guess score was** 0.04, ranged from 0 to 0.33. None of the 157 items were dropped from the test due to high guess score. Table 2, Table 3, and Table 4 show item difficulty and item discrimination scores of the 104 items for 2-3-4, 5-6-7- and 8-12 year old groups.

Analyses to determine the stopping rule of the test

Pilot data for the Receptive Vocabulary Sub-Scale were collected from less than 1000 participants, and since we needed to have responses for each item, all the items were presented to the participants. For norm data collection it was decided that the stopping rule for the test should be over inclusive. Thus "when the child fails for 8 consecutive items, he/she should be presented the rest of the items belonging to that age group and test administration should stop" was accepted as the stopping rule for the Receptive Vocabulary Sub-Scale. However, while administering a test, a target is to find the child's functioning level with the minimum number of card presentations, because when a child starts to get wrong answers and test administration continues, the child is likely to feel unsuccessful. For this reason, identification of the stopping rule based on statistical analyses appears to be important. Taking the norm data we compared the possibility of correctly responding after a certain number of failures as significantly different than .25 (since the guess possibility was .25 when there were four items and child was to choose one of them) to population rates (Newbold, Carlson ve Thorne, 2003, page 275). Until reaching the ideal stopping rule, for each ages the probability of correctly responding after 6 consecutive, 7 consecutive and 8 consecutive failures was calculated but results were unsatisfactory. Then a number of failures within a certain number of items were tested. For instance, within 10 consecutive items for 5, or 6, or 7, or 8 mistakes; within 8 consequitive items for 5, or 6 mistakes calculations were repeated. When probability results were inspected, after children made 8 mistakes within 10 consecutive items, the probability of getting the 11th item was lower than chance. Therefore for the Receptive Vocabulary Sub-Scale, within 10 consecutive items, 8 mistakes were accepted as the stopping rule.

Mean scores of norm data

IRT analysis results indicated that item difficulties were not appropriate for 13 year old youngsters. Therefore 169 youngsters aged 13 years were eliminated from the main data set of 3650 children. Than 149 children with delayed language, phonological problems, hearing impairment, learning

Table 2. Item difficulty and item discrimination scores of Receptive Vocabulary Sub-Scale for 2-3-4 years

Item	Difficulty	Discrimination	Item	Difficulty	Discrimination
2 yrs beginning			MONKEY	-0.22	1.69
TELEVISION	-1.69	0.87	TO HANG	-0.18	1.05
SNAKE	-1.34	1.02	SOAP	-0.14	1.01
DOOR	-1.24	0.95	HOSE	-0.07	1.19
CAKE	-1.23	0.98	4 yrs beginning		
FINGER	-0.97	1.28	BRACELET	-0.05	1.09
SWING	-0.87	1.05	ALONE	-0.02	1.01
LATCH	-0.81	0.89	RUBBER	0.00	1.09
BAG	-0.73	1.06	TIE	0.11	1.02
FROG	-0.61	1.02	POWER	0.15	1.03
PRETZEL	-0.57	0.99	SHEEP	0.19	1.24
3 yrs beginning			TO RUN	0.24	1.27
PILLOW	-0.48	1.13	FLY	0.29	1.29
TO KISS	-0.41	0.98	SHOES	0.37	1.04
PLATE	-0.37	1.02	BELT	0.41	1.08
ONION	-0.34	0.97	HAPPINESS	0.42	1.20
CHICKEN	-0.27	1.02	LOCK	0.46	1.29
PEARS	-0.23	1.05			

Table 3. Item difficulty and item discrimination scores of Receptive Vocabulary Sub-Scale for 5-6-7 years

Item	Difficulty	Discrimination	Item	Difficulty	Discrimination
5 yrs beginning			WHISTLE	-0.42	0.94
CHAIN	-1.65	1.02	ROCKET	-0.34	1.02
POSTMAN	-1.50	0.90	FOREST	-0.34	1.13
TO WRITE	-1.41	1.02	TELESCOPE	-0.32	0.75
DAISY	-1.31	0.83	PROPELLER	-0.31	1.04
CAGE	-1.07	0.84	WATERFALL	-0.26	1.05
DANGER	-1.06	0.93	7 yrs beginning		
SKATE	-1.04	1.00	TO DIVE IN	-0.23	1.01
VASE	-1.01	1.00	BATH TUB	-0.23	0.77
RULER	-0.98	0.82	DOCTOR	-0.02	1.18
CUP	-0.92	0.99	DIVER	-0.16	1.08
ROOF	-0.91	0.81	TEACHER	-0.12	0.89
PENALTY	-0.88	0.87	PALLET	-0.10	0.79
6 yrs beginning			SHYNESS	-0.09	0.85
DOLPHIN	-0.77	0.88	CIRCUS	-0.05	1.22
TO LOOK AT	-0.69	0.80	WALNUT	-0.05	1.06
GOAT	-0.64	0.82	ELLIPSE	-0.02	1.46
HELMET	-0.61	1.04	YOUNG TREE	-0.01	0.79
ARROW	-0.58	0.92	SAYGOODBYE	0.01	1.42
ENVELOPE	-0.55	1.04			

Table 4. Item difficulty and item discrimination scores of Receptive Vocabulary Sub-Scale for 8-12 years

Item	Difficulty	Discrimination	Item	Difficulty	Discrimination
8 yrs beginning			STRETCHER	-0.70	0.79
CYLINDER	-1.76	0.78	WAGON	-0.63	1.76
DISASTER	-1.64	1.26	FOLK DANCE	-0.57	0.76
GLORY	-1.40	1.55	COFFEE TABLE	-0.55	1.34
FACTORY	-1.34	1.07	11-12 yrs beginning		
RECTANGLE	-1.28	1.36	DAM	-0.52	0.55
KNOCK DOWN	-1.21	1.15	BREAD	-0.33	3.06
GUITAR	-1.18	2.03	PORTER	-0.25	0.68
ROPE	-1.14	1.76	STAMP	-0.18	1.76
TWITTER	-1.12	1.45	TO REPAIR	-0.08	0.78
SAILING	-1.11	1.84	TAPE MEASURE	-0.05	0.42
HALF	-1.09	1.15	HAT	-0.04	0.65
PRODUCTIVITY	-1.09	0.75	RAFT	-0.01	1.26
9-10 yrs beginning			THE OPPOSITE	0.20	2.54
RACQUET	-1.04	1.10	VIADUCT	1.43	0.58
PYRAMID	-1.01	1.64	SCOOP	1.56	0.57
LAKE	-0.92	1.70	LAMP	1.77	0.32
LONG VEHICLE	-0.77	0.69	PLOW	2.02	0.37
ISLAND	-0.73	0.56	RADIATOR	2.31	1.60
BARREL	-0.72	1.32			

disability, autism, or learning disability were further dropped from the data set, and further analyses were carried out on the data for the remaining 3332 children.

Children with fluency, articulation or resonance problems were included in the sample but 33 children whose mother tongue was not Turkish were dropped from the sample. Finally, 6 children whose raw scores were 0 when the Receptive Vocabulary Sub-Scale was revised and total item numbers were reduced to 104, were excluded from the sample. As a result means and standard scores were calculated from the remaining 3293 children's data.

Standard Scores

First, means were calculated for 1, 2, 3, 4, 6, and 12 monthly intervals to find the appropriate age intervals for standard score calculations. Results indicated that when raw scores to be transformed to standard scores; between 2;00 and 5;11 3 monthly; between 6;00 and 7;11 4 monthly; between 8;00 and 10;11 6 monthly and between 11;00 and 12;11, 12 monthly intervals would be appropriate to use.

It was assumed that in the population receptive language skills are normally distributed. Therefore, raw scores were normalized by converting them into percentile ranks in relevant age intervals. Then, percentile ranks were converted to z scores. Finally, z scores were converted to standard scores with a mean of 100 and standard deviation of 15.

Calculating Age Equivalence levels

In the use of standard tests converting raw scores to age equivalents is as useful as having standard score conversions. Thus, for each raw score age equivalents were determined. First, from 2 years to 12 years and 11 months, the sample was divided into monthly groups and for each age level the median of the raw scores was calculated. Then for monthly intervals, medians were marked on the graphic paper and a line was drawn to connect these scores. Lastly, starting from 1 for each possible raw score, corresponding age was taken as the relevant age equivalence.

Standard error of measurement

Like with other standard tests, when evaluating the results of the Receptive Vocabulary Sub-Scale, test users must consider standard errors of measurement. Standard errors of measurement are calculated by using reliability scores, and following the previous studies in the present study, split half reliabilities were used. Results indicated that for 2, 3, 4, 5, 6, 7, and 8 year old children the standard error of measurement score was

3 for 9 and 11 year old children the standard error of measurement score was 2 and for 12 year old children standard error of measurement score was 1.

Receptive Vocabulary Sub-Scale Development: Reliability and Validity Studies

METHOD

Reliability and validity studies were also carried out in parallel to norm data collection in the cities of Istanbul, Ankara, Eskişehir, Manisa, Malatya and Kahramanmaraş. Parents were given information about reliability and validity measures by the research team. Children whose parents agreed to participate in the reliability and validity studies were given the retest, Peabody, WISC-R and AGTE.

For test-retest reliability of the receptive vocabulary sub-scale, a total of 360 children aged 2-12 years were tested in 15 day intervals. Validity measures were done with 270 children aged 2-12 years by administering the Peabody, AGTE and WISC-R.

WISC-R was administered to children older than 6 years by clinical psychologists. Developmental evaluation was done by administering the Ankara Developmental Screening Inventory (AGTE) to the parents of children younger than 6 years. On the other hand, the Peabody was to all children.

RESULTS

Test-Retest

Test-retest reliability was done for each age group separately and for all ages together. Test-retest reliability varied between .70 and .94 for each age group (Table 5) and was .97 for all ages. Mean and standard deviations of raw scores for test-retest reliability was presented in Table 6.

Split-Half

Odd and even numbered items were split and the Spearman-Brown value was obtained as .99 for the whole sample. When the split-half reliability was run for each age group, Spearman-Brown value varied between .96 and .88 (Table 5).

Internal Consistency

When internal consistency coefficients were analyzed for the 104 items, Cronbach's alpha was found to be .99. Cronbach's alpha varied between .88 and .96 when it was calculated for each age group (Table 5).

Validity

While all children were tested with Peabody, WISC-R was administered to the children above 6 years and the Ankara Developmental Screening Inventory was administered to parents of children below 6 years. Correlations between the Receptive Vocabulary Sub-Scale raw scores, standard scores and WISC-R, AGTE and Peabody were calculated. Results indicated that the Receptive Vocabulary Sub-Scale standard scores were significantly related to WISC-R general, WISC-R verbal and WISC-R performance, AGTE t scores, AGTE language and cognitive subscale scores. Further, raw scores were significantly related to AGTE raw scores and AGTE language and cognitive subscale scores. However, there was no relationship between the Receptive Vocabulary Sub-Scale and Peabody scores (see Table 7).

DISCUSSION

In clinical settings, the assessment of a child's language development can be done by psychologists, speech-language therapists, audiologists and child development specialists to evaluate age appropriateness of the level of language development and to assess the progress of therapy and to make a decision for cochlear implantation. However, there has not been an original Turkish receptive vocabulary test that was developed from scratch. Researchers and clinicians in Turkey use either the Peabody test, which was adapted to Turkish in 1972 with no revisions done since then, or administer subscales of the Ankara Developmental Screening Inventory to assess language development.

In order to meet the need of a valid and reliable language test developed originally for Turkish individuals, the development of the Turkish Expressive and Receptive Language (TIFALDI) test was started in 1998. Initially, word frequency lists were formed and then 242 abstract and concrete words from different difficulty levels were selected by screening the Turkish dictionary. Finally, test booklets were prepared with one target and three distractors, with four pictures on each page drawn by a professional painter.

Pilot data were collected from 648 children age ranged between 2-13 years in villages and towns of the Ankara province. As a result of item difficulty, item discrimination and distractor analysis of pilot data, test items were reduced to 159 items: 157 for the test and 2 for the sample.

The norm data for the receptive vocabulary sub-scale of TIFALDI was collected by using a representative sample of the Turkish population determined by the Turkish Statistical Institute. Data were collected from 3755 children age ranged between 2-13 years in 61 cities nationwide. Incomplete or incorrectly collected data were eliminated from the data set during the data checking process. Finally, analysis started with data of 1760 girls, 1819 boys and 71 with no sex identification, for a total of 3650 children.

Table 5. The distribution of test-retest, split half and internal consistency reliability results by ages

Age	N	Test-retest	Split half	Internal Consistency
2	26	.94**	.94**	.94**
3	29	.85**	.95**	.95**
4	28	.92**	.96**	.96**
5	37	.78**	.96**	.96**
6	32	.81**	.95**	.95**
7	39	.70**	.94**	.93**
8	39	.76**	.94**	.94**
9	35	.87**	.92**	.90**
10	31	.74**	.92**	.91**
11	37	.74**	.88**	.89**
12	26	.76**	.89**	.88**

** Correlation is significant at the 0.01.

Table 6. The distribution of mean and standard deviation (ss) of receptive vocabulary sub-scale test-retest raw scores by age

		TEST		RETEST	
AGE	N	MEAN	SS	MEAN	SS
2	26	18.5	12	19.5	13.5
3	29	33	15	34	16
4	28	43	16	47	17
5	37	57	17	62	18
6	32	71	15	78	12
7	39	80	12	85	12
8	39	85	11	91	8
9	35	90	8	94	7
10	31	95	5	98	3
11	38	98	6	100	3
12	26	99	4	101	2

Table 7. The correlation of receptive vocabulary sub-scale raw and standard scores with WISC-R, AGTE and Peabody

Receptive Vocabulary Subscale	WISC-R			AGTE			Peabody
	General	Verbal	Performance	T score	Lang./Cog	Row score	Row score
Row score				.159	.627***	.703***	-.04
Standard score	.483 ***	.447***	.471***	.483***	.268**	.210*	.013

*** p<.000 ** p<.01 *p<.05
Wechsler Intelligence Scale for Children-Revised (WISC-R), Ankara Development Screening Inventory (AGTE), Peabody Picture Vocabulary Test (Peabody)

The TİFALDİ Receptive Vocabulary Sub-Scale IRT analyses were done by using BILOG-MG (SSI, 2002), based on the results 53 items that were extracted from the scale and some order changes were made. It was then decided that the Receptive Vocabulary Sub-Scale with 104 items is appropriate to assess receptive vocabulary skills of 2 to 12 year old Turkish children.

The TİFALDİ Receptive Vocabulary Sub-Scale reliability studies indicated that test-retest and internal consistency reliabilities were high. In addition it was found that Receptive Vocabulary Sub-Scale was significantly related to WISC-R general, WISC-R verbal and WISC-R performance, AGTE t scores, AGTE raw scores and AGTE language and cognitive sub-scale scores. However, Peabody scores were not related to either the TİFALDİ Receptive Vocabulary Sub-Scale or to WISC-R and AGTE. While results indicate that the TİFALDİ Receptive Vocabulary Sub-Scale is a reliable and a valid measure, the Peabody vocabulary scale, which was adapted to Turkish in 1972, is no longer a valid instrument to be used in the assessment of Turkish children's receptive language skills.

In conclusion, TİFALDİ Receptive Vocabulary Sub-Scale is an original, reliable and valid test, it is not adopted from another language, and norm data for this test were collected from a nationally representative sample. It is suggested that the TİFALDİ Receptive Vocabulary Sub-Scale can be used for research or in clinical settings to assess receptive vocabulary skills of 2 to12 year old Turkish children.

We plan to collect data from various clinical groups who show problems in language development to form a database. Furthermore, for the first revision of the test we plan to extend the age range, and change the cards from black and white to a colored form.

ACKNOWLEDGEMENT

Development phases of this test were funded by the Middle East Technical University Research Fund through AFP 98010402, AFP 99.01.04.04, AFP 00.01.04.03 & AFP 01-07-03-00-24 coded projects. The reliability, validity and standardization study was funded by The Scientific and Technological Research Council of Turkey (TÜBİTAK 105K151).

We would like to thank those whose valuable work made it possible to have this test today:

Project assistants Arzu Baykara, Aslı Göncü, Mehmet Akif Güzel, Tuğba Erol Pilot, Ezgi Beşikçi & Özge Sarıot. Pilot data were collected by Aslı Çakır, Ayşen Aykut, Demet Buyurgan, Deniz Tekin, Duygu Mucaoğlu, Gizem Arıkan, Gülden Elçim Üner, Hande Gürün, İlker Dalgar, İlkiz Bozkulak, Melikşah Demir, Merve Soysal, Metin Özdemir, Miri Besken, Mustafa Redzheb, Müjde Koca, Onur Sunal, Özge Orbay, Özlem Kocabaş, Pınar Önen, Rabia Ünsaldı, Selen Can, Sırma Acar, Sinem Sancaktar, Suna Türkelli, Süleyman Örikli, Şeniz Çelimli, Şirin Hacıömeroğlu, Ufuk Kılıçaslan, Yasemin Şahan, Dilek, Fulya, Kadir & Sevilay. Norm data were collected by

Açelya Konur, Ali Yıldız, Ayşe Emir, Ayşe Sarılar, Begüm Özdemir, Bilge Kaplan, Bilgen Işık, Burak Yazgan, Burcu Ergene, Burcu Subaşı, Canan Büyükaşık, Canan Karadeniz, Ceren Akdeniz, Ceren Gürbüz, Deniz Demirel, Derya Gürcan, Didem Şahin, Didem Şavran, Duygu Karabulut, Duygu Yakın, Elçin Gündoğdu, Elif Kurt, Emek Yüce, Fatih Cemil, Ferhat Satıroğlu, Gizem Sarısoy, Gonca Raslayan, Gülay Oskay, Hande Soral, İrem Metin, Marta Gurbanova, Melis Özmen, Merve Tuncel, Miray Korkmaz, Nalan Pulat, Nazlı Altın, Nermin Müftüoğlu, Nihan Kılıç, Nilgün Türkileri, Özge Tok, Özge Yaren, Özge Yılmaz, Özlem Korucuoğlu, Pınar Arslan, Saadet Bozan, Selma Yılar, Sevda Binici, Şeyda Çamlı, Sezin Andiç, Şirin Özdilek, Suzan Ceylan, Yağmur Yılmaz, Yelda Erden, Zahriye Raşitoğlu, Zeynep Gedik. IQ data were collected by clinical psychologists Dilek Sarıtaş, Gaye Zeynep Çenesiz, Mehmet Şakiroğlu & Öznur Öncül.

Data were entered and checked by Evren Etel, Ayşe Karancı & Ayça Özen. Drawings were made by Neşe Evitan. Reyhan Bilgiç & Hakan Berument helped us with the data analysis.

We would also like to thank all local officers who supported us and gave us permission for data collection from the neighborhoods. Lastly and most importantly we are in debt to all the children and families who took part in the study.

REFERENCES

Aksu-Koç A, Küntay A.C, Acarlar F et al (2011) Türkçe'de Erken Sözcük Ve Dilbilgisi Gelişimini Ölçme ve Değerlendirme Çalışması: Türkçe İletişim gelişimi Envanterleri: TİGİ-I ve TİGİ-II, TÜBİTAK 107KO58 Projesi Sonuç Raporu.

Alpas B, Akcakin M (2003) Vineland Adaptive Behavior Scales (Survey Form): adaptation, validity and reliability for infants of 0-47 months of age. Turkish J Psychol 18:57-71.

Anlar B, Yalaz K (1996) Denver II Gelişimsel Tarama Testi Türk Çocuklarına Uyarlanması ve Standardizasyonu. Hacettepe Çocuk Nörolojisi Gelişimsel Tıp Araştırmaları Grubu, Ankara.

Bates E, Marchman V, Thal D et al (1994) Developmental and stylistic variation in the composition of early vocabulary. J Child Lang 21:85 –123.

Bee H (2000) The Developing Child. 9th ed., U.S.A. Allyn and Bacon.

Bishop D (2003) Test for Reception of Grammar (Version2). UK:Harcourt Assessment.

Blank M, Rose SA, Berlin LJ (2003) Preschool Language Assessment Instrument (PLAI-2). Austin, TX: Pro-Ed.

Boysson-Bardies B, Vihman MM (1991) Adaptation to Language: Evidence from Babbling and First Words in Four Languages. Language 67: 297-319.

Dunn DM, Dunn LM (2007) Peabody Picture Vocabulary Test. 4th ed., Minneapolis. MN: NCS Pearson, Inc.

Dunn LM, Dunn DM, Styles B et al (2011) British Picture Vocabulary Scale (BPVSIII). London: GL Assessment.

Er N (1996) Çalışma Belleğinin Yapısal ve İşlemsel Kapasitesinin Faktör Analitik ve Deneysel Çalışmalarla Belirlenmesi. Unpublished Dissertation. Hacettepe University, Ankara.

Fenson L, Dale PS, Reznick JS et al (1993) The MacArthur Communicative Development Inventories. San Diego. CA: Singular.

Fenson L, Dale PS, Reznick JS et al (1994) Variability in Early Communicative Development. Monogr Soc Res Child Dev 59.

Goldfield BA, Reznick JS (1990) Early lexical acquisition: rate, content, and the vocabulary spurt preview. J Child Lang 17:171 – 83.

Hammill DD, Newcomer P (2005) Test of Language Development-Third Edition (TOLD-3) Circle Pines. MN: American Guidance Service.

Herschensohn J (2007) Language Development and Age. Cambridge University Press.

Hresko WP, Reid DK, Hammill DD (1999) The Test of Early Language Development (TELD—3). Austin, TX: Pro-Ed.

Jusczyk PW, Aslin RN (1995) Infants' detection of the sound patterns of words influent speech. Cognit Psychol 29:1–23.

Levey S, Polirstok S (2011) Language development: understanding language diversity in the classroom. U.S.A. SAGE Publications, Inc.

Martin NA, Brownell R (2011) ROWPVT-4: Receptive One-Word Picture Vocabulary Test – Fourth Edition. Pro-ed an International Publisher.

Moon C, Cooper RP, Fifer WP (1993) Two-day-olds prefer their native language. Infant Behav Dev 16:495–500.

Newbold P, Carlson W, Thorne B (2003). Statistics for Business and Economics. 5th ed.. New Jersey: Prentice Hall. p. 275.

Okalidou A, Syrika A, Beckman ME et al (2011) Adapting a receptive vocabulary test for preschool-aged Greek-speaking children. Int J Lang Commun Disord, 46:95-107. doi: 10.3109/13682821003671486.

Polka L, Werker JF (1994) Developmental changes in perception of non-native vowel contrasts. J Exp Psychol Hum Percept Perform, 20:421-35. doi:10.1037/0096-1523.20.2.421.

Püsküllüoğlu A (1997) Arkadaş Türkçe Sözlük. Arkadaş Yayınevi, Ankara.

Savaşır I, Sezgin N, Erol N (1998) Ankara Gelişim Tarama Envanteri. Publication of Turkish Psychological Association, Ankara.

Wechsler D (1992) Wechsler Çocuklar için Zeka Ölçeği (WISC-R) (Translator: I Savaşır, N Şahin). Turkish Psychological Association, Ankara 1995.

Werker JF, Desjardins RN (1995) Listening to speech in the 1st year of life: experiential influences on phoneme perception. Curr Direct Psychol Science 4:76-81.

Werker JP, Tees RC (1984) Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant Behav Dev 7:49-63.

Wiig EH, Secord W (1992) Test of Word Knowledge. San Antonio. Texas: Psychological Corporation.

Zimmerman IL, Steiner VG, Pond RE (2005). Preschool Language Scale-4. Pearson Education, Inc.

Zumbo BD (2003) Implications for translating language tests does item-level DIF manifest itself in scale-level analyses? Lang Testing 20: 136.