Anuario de Psicología/The UB Journal of Psychology
2015, vol. 45, nº 2, 147-160

An Item Response Theory analysis of the Revised Token Test in normally developing native Spanish speaking children

María Quintana[1]
Isabel Salinas González[2]
Geisa Gallardo[2]
Malcolm R. McNeil[3]
[1] *Universitat Politècnica de Catalunya*
[2] *Universidad de Guadalajara, México*
[3] *Pittsburgh University, USA*

Revised Token Test (RTT) is a proven sensitivity test for the evaluation of language comprehension. There is little evidence regarding the psychometric properties of the same in children. Similarly, many facts about the facilities of the item response theory (IRT) will not contribute to study these properties in neuropsychological tests and even less when applied to these populations. Thus, the objectives were to assess the psychometric properties of the RTT test population of children with normal development and also show the utility of the IRT models type of neuropsychological assessments. The sample consisted of a total of 250 healthy spanish-speaking children from 4 to 12 years old living in the city of Guadalajara (Mexico). This sample was divided into two groups, the older younger children (4-9 years) and (10 to 12 years). The results suggest high sensitivity of the RTT sample at different ages and also an excellent discriminant validity in both groups. In addition, the test allowed to properly classify different levels of skills related to language and other latent traits.

Keywords: *Revised Token Test, Item Response Theory, assessment, children, neuropsychology.*

Correspondence: Maria Quintana, EUNCET Bussiness School, Ctra. de Terrassa a Talamanca, km. 3, 08225 Terrassa.
Email: mquintanaaparicio@gmail.com

Análisis de Teoría de Respuesta al Ítem del Token Test Revisado en una muestra de niños nativos hispanoparlantes con desarrollo normal

La prueba Token Test Tevisado (RTT) es un test de probada sensibilidad para la evaluación de la comprensión del lenguaje. Existen escasas evidencias en relación a las propiedades psicométricas de la misma en población infantil. Del mismo modo no se aportan muchos datos acerca de las facilidades de la Teoría de Respuesta al Item (IRT) para estudiar dichas propiedades en pruebas neuropsicológicas y aún menos cuando se aplica a estas poblaciones. Así, los objetivos fueron valorar el comportamiento psicométrico de la prueba RTT en población de niños con desarrollo normal y también, mostrar la utilidad de los modelos IRT en este tipo de evaluaciones neuropsicológicas. La muestra estuvo compuesta por un total de 250 niños sanos de habla hispana entre los 4 a los 12 años de edad residentes en la ciudad de Guadalajara (México). Dicha muestra fue dividida en dos grupos, los niños más pequeños (de 4 a 9 años) y los más mayores (de 10 a 12 años). Los resultados obtenidos sugieren que RTT muestra alta senbilidad según las diferentes edades y una excelente validez discriminante también en ambos grupos. Además, la prueba permitió clasificar adecuadamente diferentes niveles de competencias relativas al lenguaje y otros rasgos latentes.

Palabras clave: *Token Test Revisado, Teoría de Respuesta al Ítem, evaluación, neuropsicología.*

Introduction

Language is a fundamental cognitive function both in the development of communication and education, as well as in social interaction. Comprehension is one of the components of language. It can be defined as the ability to decode and comprehend the meaning of linguistic messages. Furthermore, Gernsbacher and Kaschak (2006) defined comprehension as the capacity through which a linguistic input becomes meaningful. There are different instruments to assess language comprehension. Among them, the Token Test (TT), published in 1962 by De Renzi and Vignolo, is one of the most widely used. It is an easy test to administer and to understand, too. It requires the individuals to perform a series of verbal orders based on tokens with different colors and shapes which increase in complexity. Although the main objective of the TT is to evaluate verbal comprehension, it also involves other cognitive functions such as working memory, which coordinates and initiates all the processes required for efficient understanding, and is essential for the continuity needed to understand speech (Hasher & Zacks, 1988). In addition to working memory it includes the ability to understand the syntax analysis of a whole series of items, or the ability to properly ignore distracting elements (Strauss, Sherman, & Spreen, 2006).

Since the original publication of the TT there have been several versions and modifications to the test. These include the Revised Token Test (RTT; McNeil & Prescott, 1978). This test has proven sensitive to detect language deficits in patients with a diagnosis of aphasia due to a lesion in the left hemisphere in both adults and children (Gallardo, Guàrdia, Villaseñor, & McNeil, 2011; McNeil & Prescott, 1978), and adults damaged in the right hemisphere but with a diagnosis of non-aphasia (Eberwein, Pratt, McNeil, Szuminsky, & Doyle; 2007; McNeil & Prescott, 1978). The RTT has been validated and standardized in both adult and children populations (McNeil & Prescott, 1978; McNeil, Brauer, & Pratt, 1990). The most recent normative study was published by Gallardo et al. (2011) and was conducted on a Mexican pediatric population, 4-12-year-old Spanish-speaking children. Recently, the RTT has been computerized (CRTT), which allows us to increase the control over test administration and scoring. Unfortunately, the Spanish versions of this computerized test are not available yet, for which reason this paper focuses solely on the classic RTT test. As for the psychometric properties of the RTT, the abbreviated version of the five subtests has shown high test-retest reliability among adults (Park, McNeil, & Tompkins, 2000). Obtained as ICC (intraclass correlation), the value is .96 for the overall test and it ranges between .71 and .95 for the 10 subtests. The standard version has shown high internal consistency (McNeil & Prescott, 1978), higher than the shortened TT (Gallardo et al., 2011), probably due to the minor number of items of that version. These psychometric properties were obtained from the Classical Test Theory (CTT), although the RTT has also been studied from the perspective of Item Response Theory (IRT). The first of these studies was published by Willmes (1981), where the Rasch model (one parameter model) was applied, and it was concluded that the processing required for the first four subtests is fundamentally different from that required for subtest V. The second study applying IRT to the RTT was conducted by Hula, Doyle, McNeil, and Mikiloc (2006). In that paper, the authors also used the Rasch model in order to study the validity and sensitivity to change the 55-item short form of the RTT administered to individuals with a diagnosis of aphasia; the results of that study provided sufficient evidence of construct validity and content of the RTT.

Traditionally, the CTT has been the most widely used psychometric approach in neuropsychology, partly because of its simplicity and popularity (Franzen, 2011). In fact, most clinical neuropsychologists are familiar with CTT. Though a widely used theory, it is not without limitations. Among them, there is dependence on the normative group (Hambleton, Swaminthan, & Rogers, 1991) as a similar situation with the Confirmatory Factor Analysis (CFA) and non specific for IRT models and the characteristics and test scores can not be separated (Hays, Morales, & Reise, 2000). In addition, CTT assumes that the measurement error is a property of the test, and therefore, the same for all the individuals, irrespective of their score (Muñiz, 1997). CTT uses test scores as the unit of analysis.

Thus, the test scores are usually an arithmetic sum or an average of the items, or a weighted sum of component items (Franzen, 2000). As an alternative to the limitations of CTT arises the IRT.

IRT models are based on the notion that a person's performance on a particular test depends on the parametric properties of each test and the person's trait or latent ability level (Embreston & Reise, 2000). Each IRT model predicts the probability that a certain person will give a certain response to a certain item or test. People can have different levels of ability, and items can differ in many respects. IRT offers several advantages over CTT. The major challenge in neuropsychology is the invariance of the measure, in two aspects: the invariance with respect to the test, and as compared to the normative group (Muñiz, 1997). Another advantage is that we can estimate the precision with which each item and each test measure different skill levels (García-Cueto & Fidalgo, 2005), thus allowing us to have different error measures for each individual and/or skill level (Asún & Zúñiga, 2008).

It is because of these advantages that the IRT is applied in neuropsychology, taking into account that, when performing a neuropsychological examination, we are evaluating the person's performance. If we assume that not all people have the same level of ability or latent trait in a particular cognitive function, the use of IRT in neuropsychological tests and, as a consequence, in the RTT makes sense. The argument is very simple: the classical neuropsychology approach to the measurement derives from the comparison between one individual performance in a task and the normative group's average result. The RTT has not yet been studied from the perspective of IRT in a pediatric population. In this paper, our primary aim was to apply IRT to study the psychometrical properties of the RTT in normally developing Mexican children. The second aim was to point out the usefulness of IRT in neuropsychological psychometric instruments.

Method

Participants

The participants in this study were 250 healthy 4-12-year-old Spanish-speaking children randomly selected from different socioeconomic status and taking into consideration the distribution of age and sex in the population. The selection process was built by selecting different schools representing different socioeconomic status zones in Guadalajara (Jalisco, Mexico). Fifty percent of them were female. A detailed description of the recruitment procedures and the sample's characteristics has been provided by Gallardo et al. (2011) following the usual inclusion and exclusion criteria: a) monolingual; b) normal psychomotor and cognitive development, c) normal sight and hearing, and d) no history of neu-

rological problems according to teachers and parents' reports. The participants were recruited from different schools in Guadalajara, Mexico, and its metropolitan area and we obtained parental and school authorization to participate in the study in all cases.

Instruments

The RTT was administered according to the published instructions, in a Spanish version psychometrically validated in Gallardo et al. (2011). This test comprises 10 subtests varying in sentence length and complexity, each subtest having 10 homogeneous items (commands) of equal length, syntactic complexity, and vocabulary level (Eberwein, Pratt, McNeil, Szuminsky & Doyle, 2007). The RTT uses 20 tokens, including big and small circles and squares of five colors (red, blue, green, white, and black). The participants handled plastic tokens according to the standard verbal protocol. The following are examples of verbal stimuli: *Touch the big black square and the small red circle*, and *Put the small green circle to the left of the big red square*. In subtests I, III, V, VII, and IX, only large tokens are used. There are five different sentence lengths across the subtests that vary between three, four, six, and eight linguistic units to be scored in each command. Each linguistic element scored in each sentence with a 15-point multidimensional scoring system. The average of all the linguistic element scores forms a mean. Overall mean scores for each command within a subtest are averaged to obtain a mean subtest score. The maximum average score in a subtest is 15. Likewise, overall mean subtest scores for all 10 subtests are averaged to obtain an overall mean score for the entire test. The score can range between 1.00 and 15.00 and is computed in hundredths. Administration time is about 25-30 minutes. The research protocol and procedure were validated by the Bioethical Comitee of *Centro Universitario de Ciencias de la Salud* of the University of Guadalajara (Mexico).

Data Analysis

A politomic two-parameter IRT logistic (2PL) model, following Samejima's definition for Graded Response Models (1969), was used to analyze the 10 subtests and the linguistic elements of RTT. The graded response model specifies the likelihood that an examinee of a given ability will provide a response that receives a grade of X_{ij} ($X_{ij} = 0, \ldots, M$). It is important to bear in mind that we decided to study the subtests instead of each of the reactives given that the test is based on the analysis of the subtests as a global test according to the type of token each subtest uses and the task it represents. In general, IRT models are based on the study of each item's behavior. However, in this case, the peculiar role of each

item in relation to the concept of subtests makes it advisable to focus on that concept instead of using each item as an informative unit, as is common. This model predicts the probability of a correct response to any test considering the participant's ability based on two parameters: difficulty and discrimination. We selected this model because it showed a better representation than a one-parameter logistic model (it only considers the difficulty parameter), so we must include the analysis of the difficulty, but also the second parameter related to each participant's latent trait. Despite that, we estimated the fit of both models to show that the 2-parameter one was quite more satisfactory than the one-parameter one. To that end, we used the Bayesian Information Criterion (BIC), which allows us to compare models with different numbers of parameters (Schwartz, 1978). The one-parameter model yielded a BIC value = 1922.12, while the two-parameter one yielded BIC = 1412.06, clearly higher than the simple one-parameter model. From this result, we discarded the one-parameter model and focused only on the two-parameter one.

The IRT assumes that a latent trait or ability (θ) is a dimension that cannot be observed directly and is estimated based on the responses produced in a given measuring instrument. Item difficulty (parameter *b*) indicates what amount of trait is required to solve the item successfully. It describes the item's location within the ability scale. The discrimination index (parameter a) provides information about the item's discriminating power. It represents the degree to which the probability of getting an item right or wrong varies along the ability continuum, reflected by the rate of change in the probability of success as the individual's ability increases. The item response function of the 2PL model according to Samejima's proposition is defined as follows:

$$P\left(X_{ij} = x_{ij} \middle| \theta_i\right) = P^*_{Xij}(\theta_i) - P^*_{Xij+1}(\theta_i)$$

where the expression

$$P^*_{Xij}(\theta_i) = P\left(X_{ij} \geq x_{ij} \middle| \theta_i\right) = \frac{e^{D_{aj}(\theta_i - b_{xij})}}{1 + e^{D_{aj}(\theta_i - b_{xij})}}$$

contains the following terms:

P_{xij} (θ): probability of getting an item right for a value of θ
b_{xij}: item difficulty index *i*
a_j: item discrimination index *i*
D: constant (if $D = 1.7$, logistics is close to normal)

This study included children aged 4 to 12 years. In order to analyze the effect of age on performance in the RTT, it was deemed appropriate to divide the sample into age groups, in agreement with the results obtained previously through Classic Test Theory (CTT), which showed that those groups maximized the sensi-

tivity and specificity values among ages, and also because both age groups match the two great stages of development of the trait measured (Gallardo et al., 2011). Thus, the group of younger children (4-9 years old) comprised 168 children, and on the other, the older ones (10-12 years old) 84 children. This classification was based on a previous study (McNeil et al., 1990) in American children. It is also congruent with the results in a population of Mexican children (Gómez-Velázquez, González-Garrido, Zarabozo, and Amano, 2010), which showed that a substantial development change at about age 10 could account for the differential performance of the younger and older children. The item parameters for each subtest were estimated separately for the age groups. These parameters define the item characteristic curve (ICC), where each point on the curve represents the probability of getting the item right with a certain level of ability or latent trait of the individual. In addition we calculated the test information function for each age group. The same procedure was performed for the linguistic elements. In order to apply IRT polytomous models, the scores of the subtests and the linguistic elements were converted to an ordinal score according to the performance on the subtest: 1 (low), 2 (medium), and 3 (high). For this purpose, we used the percentile range. We assigned 1 (lower performance) to scores comprised in the 20th percentile or lower; 2 (medium performance) to scores within the 21-79 percentiles; and finally, high performance scores were assigned to scores at or above the 80th percentile. This recoding was based on previous literature (Peña-Casanova, Guàrdia-Olmos, Jarne-Esparcia, & Böhm, 2005; Quintana et al., 2011) showing that, through this categorization, the discrimination capacities of the subtests and indicators are kept practically intact among normative groups. In certain subtests, that conversion was not the most appropriate due to the distribution of scores and the resulting sample size of the groups. In those cases, we performed a similar conversion, thus improving the fit of the data, by using the 80th percentile instead of the 79th percentile. Table 1 (see next page) shows the conversion applied in each case.

Due to the difficulty posed by IRT models to work with missing data, and that only 1.2% of the cases had missing data, we replaced them by the Maximum Likelihood (ML) Estimation of their age group for each year, thus obtaining new observed full distributions with the same statistical properties as the original distribution but without missings (Schafer & Graham, 2002). ML techniques are preferable to any other conditional procedures based on mean imputation because they are more consistent and efficient under missing data at random conditions (Little et al., 2012).

These IRT analyses were performed using the R statistical software (v2.7) (R Development Core Team, 2011), specifically the R package ltm (Rizopoulos, 2006). For the item parameter estimation we used the ML estimation technique and latent trait by means of the Bayesian distribution of posteriori distribution. No

subsequent analyses were conducted on the differential behavior of items (DIF) as it was deemed repetitive in light of each subtest's curves.

TABLE 1. CONVERSION OF REVISED TOKEN TEST SCORES FOR YOUNGER CHILDREN AND OLDER CHILDREN.

	Younger children 4-9 years			***Older children 10-12 years***	
Subtest	*Level*	*Percentile*	*Subtest*	*Level*	*Percentile*
Subtest I, III, IV, V, VI, VII, IX, X, lingüistics elements	Low	≤20	Subtest I, V, VI, VIII, X, lingüistics elements	Low	≤20
	Middle	21-79		Middle	21-79
	High	≥80		High	≥80
Subtest II	Low	≤20	Subtest II	Low	≤20
	Middle	21-78		Middle	21-77
	High	≥79		High	≥78
Subtest VIII	Low	≤22	Subtest III	Low	≤18
	Middle	23-79		Middle	19-79
	High	≥80		High	≥80
			Subtest IV	Low	≤20
				Middle	21-77
				High	≥80
			Subtest VII	Low	≤20
				Middle	21-75
				High	≥76
			Subtest IX	Low	≤20
				Middle	21-72
				High	≥73

Results

Initially, as has been mentioned, we obtained results of the one- and two-parameter models to assess their statistical behavior. Nevertheless, the one-parameter model was later discarded as it implied too simple an approach to the problem at hand. Accordingly we opted to study the two-parameter model specifically in each age group. This way, the fit results were adequate both in the younger group (Log likelihoods = -1385.3, AIC = 2838.60 and BIC = 2944.81) and in the older group (Log likelihoods = -765.83, AIC = 1599.67 and BIC = 1681.50).

Table 2 (see next page) provides parameters estimated by the two-parameter logistic model for each subtest and each linguistic element by age group and the standard error of each estimation. As for the difficulty index, for the older children group, the subtests or linguistic elements are easier than for the younger children group (higher values and negative difficulty index).

TABLE 2. PARAMETER ESTIMATES 2-PL FOR YOUNGER CHILDREN AND OLDER CHILDREN.

	Younger children 4-9 years			***Older children 10-12 years***		
	Difficulty		*Discrimina-*	*Difficulty*		*Discrimina-*
	Extrm1	*Extrm2*	*tion*	*Extrm1*	*Extrm2*	*tion*
Subtest I	-1.308 (0,61)		4.866	-1.059 (0.41)		6.014
Subtest II	-1.694 (0.54)	1.188 (0.39)	1.290	-2.123 (0.78)	1.930 (0.43)	0.674
Subtest III	-1.486 (0.44)	1.058 (0.41)	1.714	-3.309 (1.12)	2.642 (0.63)	0.544
Subtest IV	-1.624 (0.56)	1.210 (0.37)	1.395	-2.261 (0.62)	1.835 (0.69)	0.738
Subtest V	-1.706 (0.41)	1.285 (0.33)	1.293	-1.659 (0.61)	1.464 (0.55)	1.089
Subtest VI	-1.618 (0.73)	1.204 (0.35)	1.481	-1.905 (0.69)	1.966 (0.71)	0.918
Subtest VII	-1.736 (0.69)	1.226 (0.42)	1.277	-3.357 (1.01)	2.594 (1.01)	0.450
Subtest VIII	-1.214 (0.57)	0.831 (0.28)	3.835	-2.516 (1.02)	2.510 (1.04)	0.648
Subtest IX	-1.290 (0.48)	0.842 (0.31)	1.306	-13.905 (3.99)	9.181 (2.77)	0.105
Subtest X	-1.443 (0.38)	1.057 (0.43)	1.952	-3.246 (1.21)	3.252 (1.11)	0.487
Verb I	-1.906 (0.41)	-1.903 (0.55)	5.336	-1.771 (0.89)	-1.667 (0.45)	6.383
Size I	-1.906 (0.75)	-1.778 (0.61)	4.856	-1.866 (0.77)	-1.661 (0.52)	6.949
Color I	-1.835 (0.49)	-1.761 (0.70)	4.607	-1.867 (0.59)	-1.596 (0.66)	5.118
Shape I	-1.780 (0.39)	-1.563 (0.61)	6.386	-1.870 (0.68)	-1.546 (0.72)	4.272
Verb II	-2.187 (0.48)	-0.995 (0.37)	1.164	-2.091 (0.97)	-1.033 (0.98)	1.310
Size II	-1.808 (0,88)	-1.437 (0.52)	3.736	-1.891 (0.84)	-1.420 (0.43)	2.945
Color II	-1.840 (0.87)	-1.647 (0.62)	7.186	-1.971 (0.91)	-1.503 (0.45)	2.962
Shape II	-1.785 (0.73)	-1.539 (0.67)	5.647	-1.974 (0.88)	-1.487 (0.72)	2.845
PP	-2.116 (0.93)	-2.432 (0.84)	1.207	-2.012 (0.97)	-1.358 (0.62)	2.157
LRP	-2.461 (0.78)	-0.851 (0.36)	0.861	-2.145 (1.02)	-1.351 (0.41)	1.761
AC	-2.374 (0.76)	-0.889 (0.35)	0.934	-3.224 (1.13)	-0.534 (0.18)	0.515

Note. PP = Place Preposition; LRP = Left/Right Preposition; AC = Adverbial Clause; (Standard Error for each parameter).

In terms of discrimination, most of the subtests or linguistic elements have an adequate discrimination (in IRT the values of the discrimination index are usually between 0.3 and 2.5, a greater slope in the curve indicates a good discrimination of the subtest or linguistic element, and this occurs when their values are greater than 1, according to Hambleton et al., 1991; Hays et al., 2000; Reise & Henson, 2003). In certain linguistic elements, such as size I and color I in the group of younger participants, and verb I for both groups, certain poor values are estimated as discrimination indexes.

For a better understanding of the current paper we decided to include a representative selection of the ICC obtained. Figure 1 shows the ICC in subtests V and VIII. In subtest V the tendency is similar in both groups, but the younger participants show a higher difficulty index, therefore minimal curves are located more to the right of the skill level. In the case of subtest VIII, as shown in Figure 1, the ICC values are completely different in both groups. While the younger children had a proper behavior in the subtest, good discrimination index and difficulty, in the older group the discrimination index is very low (0.544), which causes the ICC to have little slope.

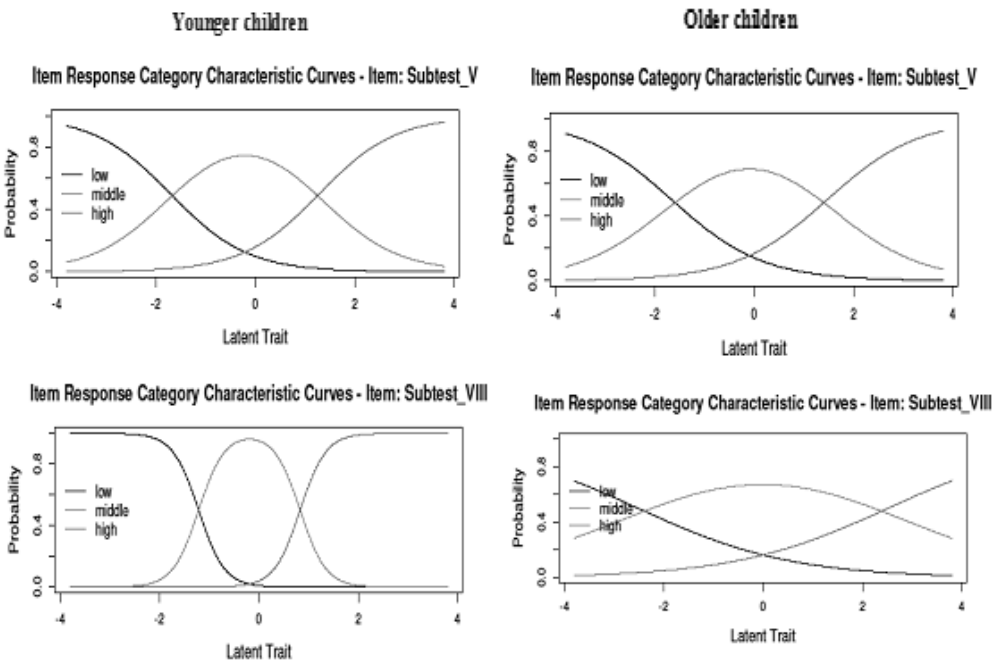

Figure 1. Item Responses Category Characteristic Curves of subtest V and subtest VIII for younger children (left column) and older children (right column).

Discussion

In the current study we applied IRT to study the RTT. As for the difficulty indexes, as expected, there are differences in the groups studied. For younger children, all the subtests or linguistic elements are harder. This indicates that a larger amount of latent traits, in this case language comprehension, is necessary to solve the subtests or linguistic elements successfully. This result is in agreement with McNeil et al. (1990), since it yields a significant change in RTT scores at age 10. However, complex understanding of language is a higher-order cognitive function that continues to develop until early adolescence (Wassenberg et al., 2008).

Onsite subtest IX in the group of older children is outside the expected range for this parameter. Therefore, it suggests that it is not working properly from the psychometric perspective (Schapira, Walker, & Sedivy, 2009) for this group of children. In the case of subtest V, the commands contain prepositional phrases that comprise six linguistic elements (e.g., Put the red circle on the blue circle). This result would lead us to the hypothesis that the younger children were able to deal with the prepositional phrase as efficiently as with the verb change (from touch to put). This result is consistent with previous studies (McNeil et al., 1990).

The estimation of the discrimination index showed that the vast majority of the RTT subtests have an adequate power of discrimination in both groups, with values greater than one in this index (Hambleton et al., 1991; Hays et al., 2000; Reise & Henson, 2003). The discrimination of an item represents the degree to which the probability of getting an item right or wrong varies along the ability continuum, which is reflected on the rate of change in the probability of success by increasing the individual's ability (Muñiz, 1997). For the group of younger children, two subtests yielded inadequate discrimination indexes: subtests I and VIII, with values above the expected range in addition to various linguistic elements (verb I, size, color, and shape I-II). This means that small changes in the skill level for these subtests in children of 4-9 years involve major changes in the probability of getting the item right. In this sense, García-Cueto (1997) proposed that items with high values in the discrimination index be eliminated from the test. This comment makes sense from a psychometric perspective, although it would be questionable in a neuropsychological functional approach. In the group of older children, we find subtests with low values of this parameter, subtests VII and IX, which indicate few discriminative subtests (Reise & Henson, 2003). Also, from the perspective of IRT, the ability of the individual himself, also known as latent trait, is studied. The best representation of this parameter is in the ICC, which reflect the individuals' degree of competence in verbal comprehension. The ICC presented shows different probabilities of getting an item right for subjects of each group studied with equal nominal standardized skills. On the one hand, children aged 10-12 years working to resolve all the RTT subtests seem to require less of an effort; whereas, on the other, children of 4-9 years require a higher level of latent trait to reach an adequate

response. The ICC has different shapes and positions. Therefore, the RTT is sensitive in discriminating different levels of ability (discriminant validity), that is, from younger children (4-9 years) to older children (10-12 years).

So far, only two studies have been published that applied IRT to the RTT (Hula et al., 2006; Willmes, 1981). In both, the sample studied comprised adults with a diagnosis of aphasia and the model used was the one-parameter logistic model, also known as Rasch Model. For his part, Willmes (1981) found significant differences among the item's difficulty estimates within each part. Furthermore, Hula et al. (2006) incorporated left-hemisphere stroke, with and without aphasia, and right-hemisphere stroke. In that study, the Rasch-derived scores are sufficiently precise to distinguish three ability levels in that sample: a) none or minimal impairment, b) mild impairment, and c) moderate to severe impairment.

In neuropsychology, the use of the CTT has a number of limitations. One of them is that the neuropsychological measure can show a non-linear distribution over the time course. In general, CTT does not use non-linear models, while IRT, based also on linear models, used generalized models, such as logistic or power models, and in some cases, non-linear models (Hambleton et al., 1991). Furthermore, it presents measurement invariance with respect to the test as compared to the normative group (Muñiz, 1997). In this respect, regarding the test's invariance, it allows us to independently make comparisons between scores and the evaluation of a skill or latent trait test in repeated measures designs. In this case, the property of invariance does not attain the relevance it presents as it is only one measurement. Because of the normative group's independence, the RTT psychometric properties described are not subordinate to the evaluation of each group's ability (García-Cueto, 1997).

However it is clear that the RTT involves other cognitive functions, in addition to the auditory comprehension, such as working memory, the ability to generate a visual image of the verbal information and the analysis of the whole within a series of elements (Lesser, 1976), and the ability to effectively ignore the automatically evoked, distracting stimulus (Winner, Connor, & Obler, 2004). Therefore, the deficit in the RTT performance can be explained by other changes than verbal comprehension.

In conclusion, from the IRT perspective, the RTT shows a high discrimination level at different ages, from younger children (4-9 years) to older children (10-12 years). Consequently, it allows us to clearly distinguish between different levels of ability or latent traits. This result, according to our considerations, is a very important improvement to the evaluation of measurement in clinical neuropsychology, especially in neurorehabilitation procedures. We must bear in mind that it is common to repeat neuropsychological measurements to estimate the performance of subjects evaluated in longitudinal designs. Fostering the idea that evaluation tests should be linked to IRT models is essential to the effects on time through measurement models have configurational invariant structure. In fact, it is more realistic to compare the performance of one individual subject to himself than comparing their performace to

the average group to evaluate their clinical response to the treatment. This a logic way to modify the clinical considerations about the measurement methodology.

This paper presents a series of limitations that should be taken into account when generalizing these results. Firtsly, the most obvious is the sample size, far from recommended values, especially when the sample was divided into age groups. This compromises both the parameter estimation system and the considerations about the impact of the results. Nevertheless, we still present them because they come from a systematically evaluated sample and one of very accurate sampling, too. In the same line of limitations, we must point out that no thorough analyses were conducted regarding the differential behavior of the items (DIF), even though it may be partially estimated based on the data presented. However, in our opinion, it would be important to improve the sensitivity and knowledge of these types of models in the most clinical and applied domains.

REFERENCES

Asún, R., & Zúñiga, C. (2008). Ventajas de los Modelos Politómicos de Teoría de Respuesta al Ítem en la Medición de Actitudes Sociales. El Análisis de un Caso. *Psykhe, 17*, 103-115.

DeRenzi, A., & Vignolo, L. (1962). Token Test: A sensitive test to detect receptive disturbances in aphasics. *Brain: A Journal of Neurology, 85*, 665-678.

Eberwein, C., Pratt, S., McNeil, M., Szuminsky, N., & Doyle, P. (2007). Auditory performance characteristics of the computerized Revised Token Test (CRTT). *Journal of Speech, Language, and Hearing Research, 50*, 865-877.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologist*. Mahwah, NJ: LEA.

Franzen, M.D. (2000). *Reliability and Validity in Neuropsychological Assessment* (2ª ed.). New York: Kluwer Academic/Plenum Publishers.

Franzen, M.D. (2011). Classical Test Theory. In J. S. Kreutzer, J. De Luca, & B. Caplan (Eds.). *Encyclopedia of Clinical Neuropsychology* (pp 586-587). New York: Springer.

Gallardo, G., Guàrdia, J., Villaseñor, T., & McNeil, M. (2011). Psychometric Data for the Revised Token Test in Normally Developing Mexican Children Ages 4-12 Years. *Archives of Clinical Neuropsychology, 26*, 225-34.

García-Cueto, E. (1997). La Teoría de Respuesta al Ítem. In G. Buela-Casal & C. Sierra (Eds.), *Manual de Evaluación Psicológica. Fundamentos, técnicas y aplicaciones* (pp. 205-219). Madrid: Siglo XXI.

García-Cueto, E., & Fidalgo, A.M. (2005). Análisis de los ítems. In J. Muñiz, A.M. Fidalgo, E. García-Cueto, R. Martínez & R. Moreno (Eds.), *Análisis de los ítems* (pp. 53-131). Madrid: Editorial La Muralla.

Gernsbacher, M. A., & Kaschak, M. P. (2006).Language comprehension. *Encyclopedia of cognitive science.* Washington: John Wiley y Sons, Ltd.

Gómez-Velázquez, F.R., González-Garrido, A.A., Zarabozo, D., & Amano, M. (2010). La velocidad de denominación de letras. El mejor predictor temprano del desarrollo lector en español. *Revista Mexicana de Investigación Educativa*, *15*, 823-847.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of the item response theory*. Beverly Hills: SAGE.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *Psychology of learning and motivation* (pp. 193-225). New York: Academic Press.

Hays, R.D., Morales, L.S., & Reise, S. (2000). Item Response Theory and Health Outcomes Measurement in the 21st Century. *Medical Care, 38*, 28-42.

Hula, W.D., Doyle, P.J., McNeil, M.R., & Mikiloc, J.M. (2006). Rasch Modeling of Revised Token Test Performance: Validity and Sensitivity to Change. *Journal of Speech, Language, and Hearing Research, 49*, 27-46.

Lesser, R. (1976). Verbal and non-verbal memory components in the Token Test. *Neuropsychologia, 14*, 79-85.

Little, R.J., Cohen, M.L., Dickersin, K., Emerson, S.S., Farrar, J.T., Neaton, J.D., … Stern, H. (2012). The design and conduct of clinical trials to limit missing data. *Statistics in Medicine*, *31*(28), 3433–3443.

McNeil, M. R., Brauer, D., & Pratt, S. (1990). A test of auditory language processing regression: Adult aphasia versus normal children ages 5-13 years. *Australian Journal of Human Communication Disorders, 18*, 21-39.

McNeil, M., & Prescott, T. (1978). *Revised Token Test.* Austin, Texas: PRO-ED, Inc.

Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Ítems*. Madrid: Ediciones Pirámide.

Park, G., McNeil, M., & Tompkins, C. (2000). Reliability of the Five-Item Revised Token Test for individuals with aphasia. *Aphasiology, 14*, 527-535.

Peña-Casanova, J., Guàrdia-Olmos, J., Jarne-Esparcia, A., & Böhm, P. (2005). Test Barcelona abreviado: desarrollo, puntuación global y validación. In J. Peña-Casanova (Ed.), *Normalidad, semiología y patología neuropsicológica. Programa Integrado de Exploración Neuropsicológica. Test Barcelona Revisado* (2ª ed.) (pp. 33-48). Barcelona: Masson.

Quintana, M., Peña-Casanova, J., Sánchez-Benavides, G., Langohr, K., Manero, R. M., Aguilar, M.,… Blesa, R. (2011). Spanish Multicenter Normative Studies (Neuronorma Project): Norms for the abbreviated Barcelona Test. *Archives of Clinical Neuropsychology, 26*, 144–157.

R Development Core Team (2011). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved November 21, 2011, from http://www.R-project.org.

Reise, S.P., & Henson, J.M. (2003). A discussion of Modern versus Traditional Psychometrics As Applied to Personality Assessment Scales. *Journal of Personality Assessment, 81*, 93-103.

Rizopoulos, D. (2006). Ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses. *Journal of Statistical Software, 17*, 1-25.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

Schafer, J.L., & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177.

Schapira, M.M., Walker, C.M., & Sedivy, S.K. (2009). Evaluating existing measures of health numeracy using item response theory. *Patient Education and Counseling, 75*, 308-314.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3ª ed.) New York: Oxford University Press.

Wassenberg , R., Hurks, P.P., Hendriksen, J.G., Feron, F.J., Meijs, C.J.C., Vles, J.S.H., & Jolles, J. (2008). Age-related improvement in complex language comprehension: results of a cross-sectional study with 361 children aged 5 to 15. *Journal of Clinical and Experimental Neuropsychology, 30*, 435-448.

Wiener, D.A., Connor, L.T., & Obler, L.K. (2004). Inhibition and auditory comprehension in Wernicke's aphasia. *Aphasiology, 18*, 599-609.

Willmes, K. (1981). A new look at the token test using probabilistic test models. *Neuropsychologia, 19 (5)*, 631-645.