DIFERENCIAS NORMATIVAS EN EL SISTEMA DE CALIFICACIÓN CUALITATIVA PARA EL TEST GESTÁLTICO DE BENDER MODIFICADO

EXPLORING THE NORMATIVE DIFFERENCES IN THE QUALITATIVE SCORING SYSTEM FOR THE BENDER GESTALT TEST MODIFIED

César Merino Soto*
Universidad de San Martín de Porres, Perú.

Recibido: 22 de enero de 2011 Aceptado: 12 de setiembre de 2011

RESUMEN

El presente estudio explora la magnitud de las diferencias en los puntajes del Sistema de Calificación Cualitativa para el Test Gestáltico de Bender Modificado, usando diferente información normativa proveniente de Perú, Estados Unidos y China. En una muestra de 324 niños(as) peruanos entre 5 y 6 años de edad, se analizaron las potenciales diferencias en la densidad, tendencia central, dispersión y clasificaciones de rendimiento visomotor. Se hallaron grandes diferencias normativas, y por lo tanto, el desempeño en los participantes se vio altamente sobreestimado o subestimado dependiendo de la norma usada. Se discute el impacto de estos resultados en la apropiada práctica evaluativa en niños.

Palabras clave: Evaluación, intercultural, Test de Bender, sistema calificación cualitativo.

ABSTRACT

This study explores the effect size in the scores of the Qualitative Scoring System for the Modified Bender Gestalt Test using different normative data from Peru, USA and China. In a sample of 324 children (as) between 5 and 6 years of age, we analyzed the potential differences in density, central tendency, dispersion and classifications of visualmotor performance. Found large normative differences, and therefore, performance in participants was highly overestimated or underestimated depending on the norm used. It discusses the impact of these results in the appropriate assessment practice in children.

Key words: Assessment, Bender Gestalt Test, Cross-Cultural, Qualitative Scoring System

Bender - Gestalt Test (TGB; Bender, 1938) continues to be, since several decades ago, one of the tests more popular, and its use has generated more than 1,000 research articles (Brannigan & Decker, 2003). To date, there are many rating systems for the TGB, and whose longevity in the history of the psychological assessment makes usually the TGB doesn’t require great presentation regarding its origin and international popularity. In the knowledge of this method for the evaluation, you must differentiate between actual test (stimuli, administration) and the rating system for scoring the reproductions. Precisely, one of the longest systems and successful methods is the System of Evolutionary Scoring by Koppitz (SES) but it was ten years after (Koppitz, 1975), in that it was refined and improved the normative information to include groups minority and more representatives sample (Makhele, Walker & Esterhuyse, 2006).

But with the advent in this decade of structural and functional changes of the TGB (for example, Brannigan & Brunner, 2002; Brannigan & Decker, 2003; Sisto, Noronha & Santos, 2006; Reynolds, 2007), would promote a resurgence of relevant research to provide psychometric data and normative comparisons. These current modifications indicate an amplitude in scaling of scores from TGB through the addition of easy and difficult items, as well as changes in the procedures of score to the designs reproduced (Brannigan & Brunner, 2002; Brannigan & Decker, 2003; Reynolds, 2007; Sisto et al., 2006).
Intercultural usefulness Potential of the TGB seems to have promoted efforts to assess the standards proposed by Koppitz (1963; 1975), and generate the own ones when the original parametric data of Koppitz was not showing appropriate «goodness of fit» with the cultural context of the normative investigations. Along with the review of the original standards, the normative research with Koppitz’s system were evaluating and confirming also the maturational hypothesis that was declaring a negative movement between age, the number and type of observed errors. The new information found in diverse cultures (for example, Chang, 1990; Ghassemzadeh 1988, Yousefi et to. 1992, Mazzeschi and Lis 1999) still support this hypothesis.

On the other hand, some studies with the TGB have shown differences beyond sampling error between the groups that show different degrees of Western acculturation. For example, in african samples (Karr, 1982; Makhele, Walker & Esterhuyse, 2006) and the middle east (Katz, Kizony & Parush, 2002; Parush, Sharoni, Hahn-Markowitz & Katz, 2000; Rosenblum, Katz, Hahn-Markowitz, mazor-Karsenty & Parush, 2000) were found differences in relation to American standards of Koppitz, (1963, 1975) as well as differences intra-groups (within the same cultures). These discrepancies appeared to be influenced by the amount of education received and the poverty level (for example, see Rajabi, 2009). Generally, the most common finding is that, compared with American standards published by Koppitz countries with lower western influence and more educational and economic inequality, the scores tend to be lower, (for example, Karr, 1982; Makhele et al., 2006). These findings converge with the first reviews of the cultural and socioeconomic and educational differences on minority groups inside the United States (Buckley, 1978).

One of the biggest distinctions between the variety of versions and modifications of TGB is the objective approach versus the global approach (Dana, Field, & Bolton, 1983); this difference relates to the methods of rating the designs reproduced, that emphasize assessments with greater or lesser degrees subjectivity in obtaining the scores. The early contributions to the line of approach have come from Pauker (1976), Keogh and Smith (1961) and the work of deHirsch (deHirsch, Jansky & Langford, 1966; Jansky & deHirsch, 1972). In this line, one of the most recent methods is the System of Qualitative Qualification (SCC, Brannigan & Brunner, 2002). Although it was influenced by Bender (1938), essentially it derived from two works in the years 60s (Keogh & Smith, 1961; deHirsch et al., 1966), who coordinated with L. Bender to develop a simplified scoring system that would allow the global assessment of designs reproduced by children. The SCC is a current refinement of the system created by Jansky and deHirsch (1972), and it uses a score for each design, that goes from 0 (random drawings, scribbles and absence of design) to 6 (exact representation of the design). This is a relatively new approach in order to qualify the reproductions obtained from the presentation of the Bender Gestalt Test Modified and it has not been reported any Hispanic research on this potential method for pre-school children. Although there are normative differences between the parametric information in China (Chan, 2001) and the American sample for the standardization of SCC (Brannigan & Brunner, 2002), to date these potential discrepancies with other groups have not been explored, such as for example, in Latin America.

The most frequent contributions on the normative differences come from the use of Koppitz’s system, in that some of these work with Hispanic population have shown inconsistent results. For example, spanish children show faster growth (Aguirre, Cortadellas & Tuset, 1988), while Mexican and Puerto Rico children (Roman & Vazquez, 1984) show a less intensive growth, as it has been replicated recently with Mexican samples (Fernandez & Tuset, 2007). In Argentina, the rules of casullo (1988) did not differ severely from the first standards of Koppitz (1963), but recent data show a superiority over both standards (Pelorosso & Etchevers, 2004). In Brazil, these patterns were also repeated (Kroeff, 1988, 1992). As it has been shown since the first normative work in other cultures, speed, the speed and slope of these changes usually happen in the first years of age, and while the age advances the intercultural differences become less intense.

Excepting the work of Chang (1990), in Peru, there have not been published works that can be identified as representative with regard to the normative comparison between Peruvian samples and standards for some version of the TGB. (The normative information of Chang 1990) was exceptional because the sample size, but it did not carry statistical comparisons between Peruvian norms and
exploring the normative differences in the qualitative scoring system for the bender gestalt test modified

standards existing for the system Koppitz (1963), as well as other validity evidences.

Therefore, as part of a line of research to introduce this version relatively new of the TGB in Spanish-speaking, and specifically in Peru, this study will assess the normative differences in the scores for a group of Peruvian children, using the SCC as a method of scoring for the modified version of TGB. Neither the System of Qualitative Qualification (SCC) of Brannigan and Brunner nor the abbreviated form of TGB have been the subject of normative research in Latin America, except in Peru, where there is a growing information psychometric published that indicates good psychometric qualities (Merino, 2009, 2010a, 2010b; Merino & Benites, 2011).

Method

Participants

The children are the population of this study in stage for the first grade of primary education from public schools in a district to the south of Metropolitan Lima, from public and private educational institutions within the same locality. The studied population belongs to the jurisdiction of the UGEL 07 (Local Educational Management Unit) according to the register at the Ministry of Education (Peru) This district is considered as a less poor (quintile 4); their indicators of poverty are less severe than other districts of Metropolitan Lima (FONCODES, 2006).

The sample of participants (n = 324; males = 196, 60.5%) will come from three public educational institutions of primary level and a preschool from the same locality. In these educational institutions, the education is uni-teaching and the teachers are female gender, and the size of each of primary reaches from 30 to 35 students. The incoming students to the first grade in public schools are not selected, because the register is free and universal (Merino, Díaz, Zapata & Benites, 2006). The children involved were not in specialized programs or ad hoc for the stimulation of skills visomotoras or language, and in general they received the amount of instruction according to rate teacher - student. The distribution of the age of our participants seems to represent usually the age range in the period of admission to the first grade of primary school in Peru (Table 1).

The functional, structural and organizational characteristics of the public schools of primary education in Peru tend to be similar, and therefore there might be assumed a reasonable equivalence in this study with respect to the children of the same public school system in urban areas.

The educational level reached by family’s head of the children evaluated is distributed as follows: primary= 17 (5.3 %), secondary= 198 (61.1 %), technical studies= 78 (24.1 %), and university= 17 (5.3 % ). In these families, mothers tend to spend more hours with the child, because they deal with the home, and eventually performed independent work; and mostly, the families of children living with other family members. Therefore, children’s homes integrate it generally more than three members, with parents of civil status of married couples or cohabitants, and belonging to the class average or less.

Instrument

Test Gestaltic Visomotor – Modified (TGB – M). This version uses only six of the original designs (A, 1, 2, 4, 6 and 8) for its application in preschool children up to the first grades of the primary level (4.5 up to 8.5 years), because these designs are most adapted for small children (Brannigan and Brunner, 2002). This version includes a scoring system of the graphics performance of the child on the six sheets submitted, called System of Qualitative Qualification (SCC, Brannigan and Brunner, 2002) of 6 points, from a punctuation of 0 (random lines, scrawled, without concept of designs) to 5 (exact representation of the design). In front of the system of Koppitz, SCC shows correlations higher with criteria of school performance (Brannigan & Brunner, 2002; Chan, 2001). This manual presents an extensive review of the psychometric findings as well as the qualification criteria of each design. In the present study, the internal consistency of the scores from SCC in our sample was $\alpha = 0.77$. 
Procedure

The data collection was carried out within a process of assessment of incoming children to two public schools, between October (2009) and March 2010. To children were administered a battery of tests of pencil-paper, exploring knowledge and pre-academic skills. Considering that its equivalence has been reported with the modality of individual administration (Koppitz, 1963; Buckley 1978; Tolor and Brannigan, 1980), the abridged version of the TGB-M was administered between 6 to 12 children per group, and in order balanced with regard to other instruments. During the evaluation of the TGB-M, the recommended standardized conditions stayed to maximise the variance relevant to construct evaluated (Bracken, 2000; McCullin, 2006) and in accordance with the guidelines for the appropriate use of test (Hambleton, 1996; International Test Commission, 2000).

The qualification of the protocols with the SCC was made by the author of the research and three undergraduate students who belong to top third; prior to obtaining the scores, we trained, monitored and evaluated the agreement between the four qualifiers, during two training sessions; the correlation between the qualifiers (model of two via random), was 0.84. On the other hand, to obtain local standardized scores, the direct scores became not linearly, with adjustment to normality, in scores T. These scores will be compared with the T-scores obtained from transforming the scores using American and Chinese standards (Brannigan & Brunner, 2002) (Chan, 2001) separately for age groups congruent with the manual.

The statistical analysis will examine the quantitative properties susceptible to show normative changes, and they will be described in the presentation of each result.

Results

The results will appear in two parts: the first one there are the statistical properties of each normative scores and their variations with regard to the age (table 1). The second part (the main analysis), compares the scores of the data Peruvians, Americans and Chinese in each of their statistical properties, and in each age level (Table 2).

Comparisons inside the normative groups. To apply an ANOVA to compare the average direct scores between the three categories of age obtained for the sample of participants (table 1), there were no statistically significant differences between them (F[2, 321] = 2.16, p = 0.11). Then, in order to verify the monotonic linear increase of average direct score between the categories of age, we applied a linear contrast a priori. It was found a linear increase in the scores between ages, L = 0.915, F(1, 321) = 4.30, p = 0.03. This result supports the construct validity of the SCC in relation to age using the direct scores.

Meanwhile, the differences in the averages of the T-scores derived from American and Chinese norms, showed different patterns. Using American standards, there were statistical differences between ages (F [2, 321] = 6.13, p = 0.002), and these differences in the T-scores average also represented a linear trend, but this time with monotonically decreasing, L = - 4.14, F (1, 321) = 7.15, p = 0.008. This indicates that the T-scores average using American standards substantially decreased as age increased. In the Table 1 it is possible to observe this tendency, under the heading «T-score - USA» for each age. On the other hand, the Chinese T standards were no statistical differences between the ages (321 (2) F = 1.80, p = 0.16), nor any substantial trend between averages; nevertheless, it is possible to recognize a slight tendency decreasing linear.

Finally, we examined the normality of scores with Shapiro Wilk’s test (1965) and Kolmogorov-Smirnov with correction Lilliefors (1967), as well as statistical of symmetry and kurtosis (table 1). A level of age exhibited approximately normal distributions, but the rest of the ages show symmetry distortions, kurtosis, or both, particularly in age 5.6 to 5.11.

The information in Table 1 also indicates that the distributional form can show important changes that are dependent on the normative information used. For example, the distortions in the symmetry tend to be similar in distributions of the direct scores of Peruvian children and in the T-scores derived from the American standard; while the use of the Chinese standards always shows differences with the previous ones. On the other hand, the empirical distributions are more similar to the theoretical normal distribution in the ages 5.0 – 5.5, and 6.0 – 6.5. In the Figure 1, shows graphically the form of the distributions of scores T.
## Exploring the Normative Differences in the Qualitative Scoring System for the Bender Gestalt Test Modified

### Table 1

<table>
<thead>
<tr>
<th></th>
<th>Age 5.0 – 5.5</th>
<th>Age 5.6 – 5.11</th>
<th>Age 6.0 – 6.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n = 51)</td>
<td>(n = 169)</td>
<td>(n = 104)</td>
</tr>
<tr>
<td><strong>Scores T</strong></td>
<td>USA</td>
<td>China</td>
<td>USA</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>18.745</td>
<td>62.37</td>
<td>49.05</td>
</tr>
<tr>
<td><strong>Of</strong></td>
<td>3.637</td>
<td>10.55</td>
<td>23.5</td>
</tr>
<tr>
<td><strong>Min. – Máx.</strong></td>
<td>11 - 28</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Symetri</strong></td>
<td>0.11</td>
<td>0.14</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>(0.32)</td>
<td>(0.44)</td>
<td>(0.105)</td>
</tr>
<tr>
<td><strong>Kurtosis</strong></td>
<td>0.14</td>
<td>0.07</td>
<td>-0.22</td>
</tr>
<tr>
<td></td>
<td>(0.20)</td>
<td>(0.11)</td>
<td>(-0.33)</td>
</tr>
<tr>
<td><strong>Shapiro-Wilk Test</strong></td>
<td>0.97</td>
<td>0.98</td>
<td>0.976</td>
</tr>
<tr>
<td></td>
<td>0.091</td>
<td>0.091</td>
<td>0.086</td>
</tr>
<tr>
<td><strong>Kolmogorov-Smirnov (KS) Test</strong></td>
<td>0.091</td>
<td>0.091</td>
<td>0.086</td>
</tr>
</tbody>
</table>

**Comparison between normative scores**

Differences in density. To evaluate the empirical distributions, will be used graphics kernel and the Kolmogorov-Smirnov Test for two samples (KS, Smirnov, 1939), which makes a bivariate analysis of the null hypothesis of equality of the empirical distributions. This statistical test is sensitive to the differences due to the location, dispersion and/ or form of the the empirical accumulative function in the scores of the groups (Pett, 1997). Table 2 shows that differences in the density occur in all age levels, and this is generally higher among Peruvian and American parameters. By means of graphics kernel, Figure 1 presents the distributional differences in each age level. It is noted that the origin of the inequity distribucionales happens in the dispersion and location, which causes differences of varied magnitude in the kurtosis and symmetry, as we discussed above (see Table 1).

**Differences in central tendency.** In Table 2, there are the results of comparisons between the normative groups inside each age group, with regard to average values. We applied the Student’s t test for dependent samples, but due to the distributional asymmetry of scores, it was used the bootstrap procedure to obtain the values p to evaluate the statistical significance. To mitigate the effect of the distributional non - normality, It is also derived robust versions (Algina, Keselman & Penfield, 2005; Hogarty & Kromrey, 1999) of magnitude’s estimation of the effect based on standardized differences, of Cohen (Coe & Merino, 2003), through a macro ad hoc for the program SAS (Kromrey and Coughlin, 2007) that will use the averages cut away in the calculation (d Cohentrim). The analyses show that, except in the comparison Peru - China in the first and third age level, the rest of the differences t-Student test were statistically significant. The differences between the Peruvian and American standards tended to decrease as the increase of the age; and the magnitude of these differences was very high, because these were greater than 3 deviations standards of the differences between averages. The differences between the averages between Peru and China also increased with age, being more well-known in the third age level.

**Differences in variance.** We compared the variances between each pair of SD through a test of differences for variances related (Pitman, 1939; Howell, 1997). The variability in T - scores extracted from Chinese and American standards showed statistically significant differences in relation to the variability of scores based on own statisticians of the Peruvian sample. The SD of T - scores for Peruvian sample in three age levels were 9.69, 9.86 and 9.81, respectively. The differences in the variability of the scores,
in general, were larger among the Peruvian sample and the Chinese standards, and the tendency of these differences was that decreased with age, but always the difference was almost twice the size of the dispersal of the Chinese standards. For example, according to the reason for the dispersal of the Peruvian scores (SD= 9.96) in the age 5.0 to 5.5 in front of the dispersion of the Chinese sample (SD= 23.5, Table 1), this was almost half of the last one; in other words, twice the dispersal of Peruvian standards. On the other hand, the dispersal of American standards to Peruvian standards was lower, but still statistically significant according to T test of Pittman (Pitman, 1939; Howell, 1997).

Differences in performance classifications. To evaluate the concordance of performance estimate using different standards, T-scores were transformed into qualitative ratings of performance. To obtain these classifications of performance in qualitative levels, we categorized the performance of children in accordance with different standards used. It was chosen to classify the performance on 7 levels of T-scores: less than 29, 30 to 35, 36 to 42, 43 to 55, 56 to 62, 63 to 69 and more than 70. This separation is consistent with the common forms of splitting the performance in several sections to differentiate better the performance between 2 standard deviations around the average (Reynolds, 2007).

In all ages, the percentage of agreement in the scores classified according to the previous paragraph was below 50%, and mostly around 30% of agreement. The poorest
agreement occurred in the first age level, which hardly differed from zero. These estimates of the agreement may be biased by the random agreement; therefore, we use the Kappa coefficient (Cohen, 1960) to make a better estimate of the concordance. Kappa’s coefficients revealed a worse performance classificatory between the Peruvian and non-Peruvian standards. Although this agreement tended to rise with the age level, the found agreement levels are considered to be conditionally poor (Cicchetti, 1994). On the other hand, using the model of two-way to Intraclass correlations (ICC; Shrout and Fleiss’, 1979) and like minimal limit the value of 0.70 to separate to the qualifications with acceptable precision (Cicchetti, 1994), the results tended to be slightly low. The agreement, nevertheless, tended to raise in the late age levels, and the magnitude of the ICC in the last age level might be considered to be satisfactory thinking that it will overcome to the comparison parameter.

**Discussion**

This research focused on evaluating the normative information of a new system of qualification for the TGB-M, in Peruvian children. The finding indicate that the use of foreign standards with regard to the SCC for the TGB-M has shown a degree of parametric variability which seriously compromises the practice to apply it in educational and administrative decisions in Hispanic children, specifically in Peru. Although it requires a much larger sample size to establish statistical more stable to compare them in front of USA and Chinese existing standards, we have quantitative and comparable evidence to assert that the differences are not sampling error and that they go beyond the simple averages difference. This study provides data for one of the new versions of the TGB applicable to children and can be very useful in a battery of development’s assessment or detect future problems in school performance.

As there is no other normative information available to date on the use of SCC in Hispanic samples at the global level, then we cannot know the degree in which the Peruvian normative information presented here may differ from other Hispanic groups in their own country of origin, or groups of Hispanic immigrants USA.

### Table 2

**Comparisons between the empirical scores and American and Chinese standards derived from manual (Brannigan and Brunner, 2002)**

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Age 5.0 – 5.5 (n = 51)</th>
<th>Age 5.6 – 5.11 (n= 169)</th>
<th>Age 6.0 – 6.5 (n = 104)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Perú vs USA</td>
<td>Perú vs China</td>
<td>Perú vs USA</td>
</tr>
<tr>
<td>A. Differences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Distribution (KS)</td>
<td>0.49**</td>
<td>0.46**</td>
<td>0.26**</td>
</tr>
<tr>
<td>2. Central Tendency</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t of Student</td>
<td>-70.82**</td>
<td>-31.28**</td>
<td>-28.01</td>
</tr>
<tr>
<td>Average Dif.</td>
<td>-12.33</td>
<td>-11.64</td>
<td>-6.44</td>
</tr>
<tr>
<td>d Cohen</td>
<td>-9.08</td>
<td>-5.57</td>
<td>-3.37</td>
</tr>
<tr>
<td>3. Variability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratio</td>
<td>1.089</td>
<td>1.43</td>
<td>1.16</td>
</tr>
<tr>
<td>T Pittman</td>
<td>3.34**</td>
<td>18.48**</td>
<td>4.68**</td>
</tr>
<tr>
<td>B. Agreement</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC (2, 2)</td>
<td>0.57</td>
<td>0.63</td>
<td>0.82</td>
</tr>
<tr>
<td>% of agreement</td>
<td>1.96%</td>
<td>22.48%</td>
<td>44.23%</td>
</tr>
<tr>
<td>Kappa</td>
<td>-0.13</td>
<td>0.043</td>
<td>0.19**</td>
</tr>
<tr>
<td>r Pearson</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
</tr>
</tbody>
</table>

****: p < 0.01. *: KS: Kolmogorov-Smirnov Test. +: ICC (2, 2): Intraclass correlation, Two tracks random
Although the choice of the sample for this study is a limitation to generalize our results, we can consider that the magnitude of the low probability of committing a Type I mistake found in our research, as well as the magnitude of the differences in several statistical characteristics (distribution, central tendency, variability and the agreement in the classification of performance), have been large enough to assume that these differences may be the result of systematic influences. These possible influences can happen in the social and educational experiences of children involved, which is a commendable and coherent explanation with other investigations that have found differences between cultural groups (Chang, 2001; Fernandez and Tuset, 2007; Makhele et al., 2006). The differences found cannot be ignored in the decisions on the choice of the most appropriate standards in the exercise of the use of the SCC, and even the use of other tests of performance visual-motor. As other aspects of the child development, the impact of the culture depends on the quality and quantity of received stimuli, as well as on the interactions in which these happen, and must be considered to be a perspective emic-ethical when it cover the differences of these conditions on the validity of a construct (Kagitcibasi, 2004).

One of the results was that the T-scores average using American standards systematically overestimated the performance in the BGT in the first two age levels, and approached the central parameter of 50 in the last age level. This overestimate meant that in the first semester of the age of 6, the distortion is 3 standard deviations under the average, while in the whole status of the age 5, the overestimate changed between 9 and 5 standardized units T. In all ages, the discrepancy between the children of the American standard and Peruvian children in our study clearly favors the latter. The opposite happened using the Chinese standards, in that the T-scores tend to underestimate the performance visual-motor; this happened in the last two age levels and slightly below 50 in the first age level. This seems to indicate that the differences between the visual-motor performance of Chinese children and Peruvian children involved are accentuated more when they come to age of 6, and that the discrepancies usually favor Chinese children in the age range evaluated, because the use of Chinese norms produce a consistent underestimation of the visual-motor skill in Peruvian children of our sample.

As well as one of the limitations of the study is the sample size of some evaluated age status, the sample only comes from an urban district in metropolitan Lima. There is an open question on the generalisability of these findings, but that future research will be able to respond on this point.

Our results point to conclude various aspects. In each age group, the descriptive distortion is much higher using American standards that using the Chinese standards; but in both situations, the choice of any of these rules is inappropriate because it located normatively to the child assessed in severely wrong positions. Where the difference in the location was lower and corresponding to a moderate discrepancy (for example in the first two levels of age in the comparison Peru - China), it might seem that to use the foreign norm might lead to only mild or moderate distortions in the description of the assessed person. But the impact may be really severe if we consider the full range of the possible scores because in any of the levels of punctuation the differences can be major than those that occur in other sectors of scores’ scale. For example, the effect of the great dissimilarity in the dispersion means that the scores more far away from the average will tend to be highly over-estimated or undervalued. And an important practical consequence of this distortion is the generation of false positive or false negative, with regard to the identification of skills related to problems in learning. The information presented in the study not only indicates that there are statistically significant differences between groups, but also an estimate of the magnitude of these differences. In the studies reviewed in the introduction to this article, estimates of this kind are not reported, and it might not evaluate if the differences have also a practical value and not just statistical. If the reader asks how much could be the difference between the scores obtained by a child in the modified version of the Bender, using own standards, the American or Chinese norm, then with security we could say that there would be a severe difference, and a consequent distortion in the diagnosis. However, this distortion would change with the age.

In practice, the use of foreign standards brings problems not only diagnoses and descriptive, but also ethical. If a large group of children of different ages, is rated with a test for which there are available foreign norms, and the statistical description of the groups will be done with such rules, which
may anticipate is that there would be a reasonable uncertainty on the appropriate description of the children and groups evaluated. In other words, children could have scores that mean overestimation or underestimation and an appropriate estimate of their performance. If this problem is compounded by the effect of the measurement error, then we will have an estimation of the level of skill that would not be useful in any applied circumstance. Ethical guidelines on a use responsible and informed about the tests from several sources (American Educational Research Association, American Psychological Association and National Council on Measurement in Education, 1999; International Test Commission, 2000) require compliance with the psychometric qualities of measuring instruments, but they emphasize even more the responsibility on the decisions in the election of these instruments and consequences of inferences derived from its uses. And to use foreign norms without an assessment of their adequacy in the context of its use, it may not be less than an example of inappropriate practice.

This study corroborates the problematic of this situation in the applied area and research, because the estimations of performance obtained by inappropriate norms create an inaccurate image of the behavioral status of a child or evaluated group. But, since a methodological framework, this investigation could reveal the impact of the normative differences in several sources of parametric information. Therefore, you can consider that an analysis expanded toward such sources can give a better figure of the quantitative changes in the normative differences of the instruments in study of psychometric adaptation. And finally, if the modifications to the TGB are concurrent with the clinical challenges and the psychometric methods that each modification produces (Dana, Field and Bolton, 1983), then the normative information must be a space of evaluation as accurately as possible. In this situation, the study presented here can make a normative contribution with regard to the new abbreviated version of the TGB, but also methodological considering the aspects evaluated sequentially.

References


* sikayax@yahoo.com.ar

ISSN: 1729-4827


