WCPCG-2010

A comparison of clinical and nonclinical groups of children on the Bender-Gestalt and Draw a Person Tests

Serap Özer a *

aDoğuş University, Acıbadem, İstanbul, Turkey

Received January 3, 2010; revised February 9, 2010; accepted March 4, 2010

Abstract

The present study compared a clinical and a control sample of Turkish children on the Bender Gestalt and Draw A Person tests. 44 of the children from a clinic sample were compared to 44 children from a matched nonclinical school sample. The tests were scored according to the Koppitz criteria. ANOVA comparisons showed differences on the Bender Gestalt test, and the HFD. The two groups did not differ on the number of Emotional Indicators. Correct classification of the children in the clinical groups based on the presence of Emotional Indicators was accurate only in 56% of the cases.

Keywords: Test, Psychological, children, drawings, Draw a Person, Bender Gestalt.

1. Introduction

Both the Bender Gestalt test and Draw a Person test not only have a historical place in Clinical Psychology but also maintain their popularity today. Both of these tests are among the most frequently utilized instruments as reported in surveys of clinical psychologists (Camara, Nathan & Puente, 2000; Cashel, 2002). Especially when psychologists working with children are surveyed (Camara et al., 2000; Cashel, 2002) these tests are ranked in the top 5 most frequently utilized tests.

1.1. Draw A Person

There are two main areas where human drawings are used with children: The first is the attempt to do a global prediction of the child’s cognitive development (Brown 1990; Cherney, Seiwert, Dickey, & Flichtbeil, 2006; Koppitz, 1968; Naglieri, 1988). The second area of utilization is to evaluate the emotional make up or the emotional state of the child (Catte & Cox, 1999; Matto, 2002; Tharinger & Stark 1990). In emotional evaluations the presence of signs that a theoretical framework (which has traditionally been an analytical framework) has deemed important has been used as foundation for clinical interpretation (Hammer, 1958; Machover, 1949; Yavuzer, 1992). However, these types of interpretations have not received empirical support (Motta, Little, & Tobin, 1993; Garb, Wood, Lilienfeld, & Nezworski, 2002). For example no relationship has been shown between anxiety, depression and

* Serap Özer. Tel.: +90 216 544 5555; fax: +90 216 544 5533
E-mail address: sozer@dogus.edu.tr

© 2010 Elsevier Ltd. Open access under CC BY-NC-ND license.

doi:10.1016/j.sbspro.2010.07.121
emotional disturbance and signs proposed by systems of drawing interpretation (Joiner, Schmidt & Barnett 1996, Motta et al. 1993). Systems have been developed (Koppitz, 1968; Naglieri, McNeish, & Bardos, 1991) where an attempt has been made to look at total number of indicators and attempt to relate these to differentiation of emotional or behavioral disturbance (Catte & Cox 1999; Matteo, 2002; Tharinger & Stark, 1990). For example, in Koppitz’s (1968) system it is suggested that 3 or more emotional indicators need to be present to determine emotional problems. Catte and Cox (1999) who used this system in their research, were able to show that a clinical group of children showed a higher total number of emotional indicators in their drawings than a control group.

The Koppitz (1968) system has received empirical support in clinical comparisons (Catte & Cox, 1999; Matteo 2002) and has been frequently utilized in different cultures (Catte & Cox 1999; DiCarlo, Gibbons, Kaminsky, Wright, & Stiles, 2000; Vedder, Van De Vijfeijken, & Kook, 2000). The Koppitz system has 30 developmental items (such as correct number of fingers, presence of legs in two dimensions etc.). It is possible to obtain a score indicating the child’s developmental cognitive maturity based on the frequencies of these items.

In defining Emotional Indicators, signs that were theoretically expected to reflect emotional problems were selected and comparisons of clinical and control groups were undertaken. In the final selection of Emotional Indicators 3 main criteria were used: 1) an item needs to be observed with unusual (less than %15) frequency in the normative group; 2) the item should not be affected by age or cognitive development; and finally 3) clinical validity, that is differences between a clinical and control group should be demonstrated. There are 30 Emotional Indicators defined with these criteria in the Koppitz (1968) system.

Cultural differences are also important in determining Emotional Indicators. Catte and Cox (1999) in a study with English children have shown that only 16 of the 30 Koppitz Emotional Indicators fulfilled the criteria for being observed with less than %15 frequency and being unaffected by age. Although there are studies that utilize Draw A Person as a measure in Turkey (Kağıtçıbaşı, Sunar, & Bekman, 2001), there are no normative studies on DAP in Turkey other than an earlier attempt utilizing the Goodenough Harris system (Uçman, 1972). A study with a Turkish sample (Özer, 2009) has shown that Koppitz EI’s are observed with frequencies different from U.S. norms.

1.2. Bender Gestalt

The second test utilized in this study, the Bender Gestalt (BG) test, is probably one of the most historic tests in psychology. It was originally proposed by Bender (1938) to measure organic integrity. Later systems (Brannigan, 2002; Koppitz, 1963, 1975) have focused on the number of errors made by subjects in copying the designs. Participants are presented with 9 figures and are asked to copy them.

The Koppitz (1963) system was selected for this study, instead of the more recent Brannigan (2002) system based on two reasons. First, when we review Turkish psychology literature instances of the use of the Koppitz system is observed (Antikacoğlu, 1987; Karaman, Türkbay, Tümer, & Gökçe, 2006; Kiriş & Karakaş, 2004, Seven, Soysal, & Cinaz, 2002). Second, there are many studies utilizing the Koppitz system in non-Anglo-Saxon cultures. (Chan, 2000; Özer, 2007; Robin & Shea, 1983; Viljoen, Levet, Tredoux, & Anderson, 1994).

It appears important to investigate the differences between control and clinical groups on these tests. Although there are some publications in Turkey (Yavuzer, 1992), on the clinical interpretations of children’s drawings there are no empirical studies (Saylı, 2004). There are studies (e.g. Kağıtçıbaşı et al., 2001) that have utilized human drawings as a measure of emotional maturity but none that have looked at clinical screening. No studies looking at the effectiveness of BG as a screening device differentiating clinical and control groups have been conducted. The purpose of the present study is to investigate the performance of a clinical and control group of children on these two frequently utilized tests and look at the effectiveness of these instruments as a screening instrument.

2. Method

2.1. Participants

88 children ranging in ages between 5 years 11 months and 11 years participated in this study. 44 of these children were in the Clinical group. These were children who have applied for psychological services, in a free university clinic or a private clinic for behavioral and emotional problems. Since the aim of the study was to look at a global effectiveness of these instruments in separating clinical and nonclinical groups, the various diagnostic labels
in the clinical group were not taken into account but rather the fact that they required psychological services was considered to be the sole criteria. The comparison group consisted of protocols selected from a sample of children attending public or private schools who were administered the tests during routine school evaluations. None of these children were receiving or were referred for psychological services. 44 children matched for age, sex and school types were selected. 64 of the participants were male, and 24 were female. 52 of the children were attending private school, while 36 were attending public school.

2.2. Procedure

All tests were administered individually and in random order. The administrations of the tests in schools were undertaken by school counselors or clinical psychology graduate students who had received the same training in administration. In the clinical group the children were administered the tests by the investigator or by Clinical psychology graduate students trained by the investigator. For the HFD, the children were provided with one A4 paper and were instructed to “draw a person, a whole person, and not a stick figure” (Koppitz, 1968, p. 6).

The instructions for the administration of the Bender Gestalt test were from Koppitz’s (1975) procedures. That is, there were no time limitations, the children were asked “to draw the figures as they see them”. The design card was not removed until the drawing was completed. The children were allowed to manipulate the cards and their drawing papers. The orientation of the cards and the paper while each design was copied was noted. For both administration children were provided with erasers and there were no time limitations.

2.3. Measures

Human Figure Drawings (HFD) The HFD. Koppitz (1968) developed a procedure for assessing the presence or absence of various items in Human Figure Drawings (HFD) that could be expected to appear in the performances of 5- to 12-year-old children. There are 30 details that can be included by the children in their drawings. Reliability studies (Koppitz, 1968; Rae & Hyland, 2001) indicate that the inter-rater reliability commonly exceeds 0.90.

For scoring Koppitz (1968) criteria were utilized. The presences of any of the 30 details that can be drawn by the children were noted. A total score was obtained for each child through the addition of the total number of details drawn by them. Then, all the drawings were scored for the presence of EI’s specified by Koppitz (1968) for each age level. 20 of the drawings were randomly selected and were scored by two graduate level clinical psychology students trained in the Koppitz method. The Pearson correlations of the total number of details scored by the primary investigator, to the scores obtained by the other judges were satisfactory ($r= .82$ and $r= .87$). For the reliability measurement of EI scoring, the correlations between the investigators total number of EI scores and the two judges were calculated and found to be $r = .91$ and $r = .93$. Since an earlier study (Özer, 2009) has shown that some EI’s specified by Koppitz (1968) may appear with more than 15% frequency in a Turkish sample or the absence of some items at certain age levels is observed with a high level of frequency, a modified list of EI’s excluding 12 of Koppitz’s EI’s were utilized. Comparisons between the clinical and comparison groups were also undertaken for these Modified EI’s.

Bender Gestalt (BG) The Bender-Gestalt test consists of nine 4 in. X 6 in. cards, each displaying a unique figure. The individual is asked to draw each figure as he or she observes it. After testing is completed, the results are scored based on the presence of any of the 30 possible errors defined clearly in the Koppitz (1975) Developmental Scoring System. These errors are defined as distortion of shape, rotation of whole design or part of it, failure to integrate the design and perseveration. Since errors are scored, a higher score indicates a poorer performance. Inter-rater reliabilities for the system are reported to range between .79 and .97 (Sattler, 2002). Test retest reliability ranges between .50 and .84 (Koppitz, 1975).

The protocols were scored by the researcher following Koppitz’s (1975) criteria. The test protocols were scored for the presence/absence of the 30 error definitions as defined by the Developmental Scoring System (Koppitz, 1975). A total error score was obtained for each child. Twenty of the Bender-Gestalt protocols were randomly selected and scored by two independent judges who were M.A. level Clinical Psychologists who had received training in the Koppitz (1975) method. Inter-rater reliabilities were calculated based on correlations between the investigators scoring and each judges score ($r=.80$ and $r=.89$). These were consistent with the reliabilities reported by other studies using this system (Koppitz, 1975; Mazzechi & Lis, 1999).
3. Results

The scores obtained by the children in the clinical and comparison groups on the BG and HFD are presented in Table 1. When ANOVA comparisons were completed it was observed that the clinical group of children obtained higher error scores on the BG ($F= 4.60, p < .05$) and had a fewer number of details in their HFD’s ($F= 4.60, p < .05$).

<table>
<thead>
<tr>
<th>Groups</th>
<th>Clinical (n=44)</th>
<th>Comparison (n=45)</th>
<th>Total (n=89)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BG Total Error</td>
<td>Mean 4.75</td>
<td>Mean 3.18</td>
<td>Mean 3.96</td>
</tr>
<tr>
<td></td>
<td>SD 3.86</td>
<td>SD 3.35</td>
<td>SD 3.68</td>
</tr>
<tr>
<td>DAP Total Detail</td>
<td>Mean 14.23</td>
<td>Mean 16.78</td>
<td>Mean 15.51</td>
</tr>
<tr>
<td></td>
<td>SD 6.29</td>
<td>SD 4.84</td>
<td>SD 5.72</td>
</tr>
<tr>
<td>DAP Modified EI</td>
<td>Mean 1.58</td>
<td>Mean 1.11</td>
<td>Mean 1.34</td>
</tr>
<tr>
<td></td>
<td>SD 1.46</td>
<td>SD 1.15</td>
<td>SD 1.32</td>
</tr>
<tr>
<td>DAP EI</td>
<td>Mean 2.02</td>
<td>Mean 1.58</td>
<td>Mean 1.79</td>
</tr>
<tr>
<td></td>
<td>SD 1.78</td>
<td>SD 1.63</td>
<td>SD 1.71</td>
</tr>
</tbody>
</table>

When we look at EI’s scored on HFD, contrary to expectation no significant differences were observed between the clinical and comparison groups ($F = 1.51, p > .05$). The number of Modified EI’s drawn by the children also did not produce significant differences between the groups ($F = 2.81, p > .05$).

In clinical practice rather than looking at EI’s specifically it is recommended that a cut-off point of 3 or more EI’s be considered as a global indicator of psychopathology (Koppitz, 1968). In order to reflect this practice the children were grouped according to their number of EI’s and their actual placement in Clinical or Comparison groups were investigated (Table 2).

<table>
<thead>
<tr>
<th>Groups</th>
<th>3 or more EI’s (n=30)</th>
<th>2 or less EI’s (n=59)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical</td>
<td>18 (%41 correct placement)</td>
<td>26 (%59 false negative)</td>
</tr>
<tr>
<td>Comparison</td>
<td>12 (%26 false positive)</td>
<td>33 (%73 correct placement)</td>
</tr>
</tbody>
</table>

It can be seen that 18 of the children who drew 3 or more EI’s came from the clinical group, while 12 were from the comparison group. 26 of the children from the clinical group had DAP’s with 2 or less EI’s, thus placing them in the wrong group based on the 3 or more EI criteria. That is, the “false negative” error rate for this group was 59%. In other words while the use of number of EI’s in group placement was more successful in correctly placing the children in the Comparison group, more than 50% of the children in the Clinical group had DAP’s with less than two EI’s which would incorrectly classify them as “Comparison” group.

4. Discussion

When the performances of the clinical and comparison group of children on the Bender Gestalt and Draw A Person tests were undertaken it was observed that for both tests there were significant differences between the groups on the developmental indices provided by the Koppitz (1963, 1968, 1975).
The Developmental Scores on the Bender Gestalt test showed a difference between the clinical and comparison groups. It was observed that the children in the comparison group showed a higher level of visual spatial maturity. The use of Draw A Person as a screening device in distinguishing clinical and nonclinical groups receives partial support. The two groups differed on the developmental indices on this test (the total number of age appropriate detail used in drawings) however the same type of differences were not observed on emotional indices. The fact that the children in the comparison group not only showed more maturity on the Bender Gestalt test but also on the DAP indicates that these children were more mature cognitively than the clinical group of children. While there are some studies looking at the relationship of general cognitive ability (Abell, Von Briesen, & Watz, 1996, Aikman, Belter, & Finch, 1992) or school readiness (Szasz, Baade, & Paskewicz, 1980) and children’s drawings, studies looking at differentiating clinical or groups based on developmental indices are not present. The present findings are important in this respect. Although clinical practice assumes the differences between clinical and nonclinical groups to be based on emotional differences, it was shown that cognitive maturity can also be an important differentiator.

When we look at the findings from the DAP test it can be observed that many interpretations based on theoretical formulations (Koppitz, 1968; Yavuzer, 1992) is not supported by this study. There were no differences in the number of Emotional Indicators in the drawings of the children from clinical or nonclinical groups. There have been studies that were able to demonstrate such differences in other cultures (Catte & Cox, 1999). The reason for the lack of differences in this study may arise from the fact that there has been no standardization study done to define emotional indicators for Turkish children. It was shown in another study (Özer, 2009) that signs such as “missing hands” which are theoretically important emotional indicators (Koppitz, 1975) were observed with a “not unusual” (more than 15%) frequency in a Turkish sample. The continued use of DAP in Turkey is dependent on a normative study which will define frequencies of EI’s.

The use of these two tests as a clinical screening instrument is not warranted based on the results of this study. It is the aim of screening devices to increase the rate of “false positives”, that is allow nonclinical individuals to be screened in, and allow further more sophisticated assessment devices to complete the differentiation. The results of the present study indicate that “false negative” errors are more frequent in the Draw A Person test. That is, the use of this test as a screening device will mean the misclassification of a large number of children with psychological problems as nonproblematic. This would be an important limitation for this test. It should be noted that the present study utilized a small, convenience sample; hence the results can only be regarded as exploratory. However, the need for larger, controlled comparison of this test in a Turkish sample is warranted. The emotional indicators on the DAP should not be regarded as confirmatory in the diagnostic process but rather as contributing data in the general assessment process. As Sayil (2004) has indicated it is very important to limit the contribution of drawing data in the assessment procedure. The developmental indices obtained from the DAP, especially in conjunction with Bender Gestalt results may be an indicator of spatial-motor maturity and serve a purpose as a screening device.

In summary, it is important for clinicians doing assessment with children to be aware of the limitations of these two easy to use, economic tests. These may be utilized as “ice breakers” in communicating with children, may provide information about the general level of spatial-motor maturity, however, interpretations on emotional structure and diagnosis are not warranted.

References


