Scandinavian Journal of Psychology, 2008, 49, 239–246 DOI: 10.1111/j.1467-9450.2008.00637.x

Cognition and Neurosciences

A comparison of normative data for the Trail Making Test from several countries: Equivalence of norms and considerations for interpretation

ALBERTO L. FERNÁNDEZ[1,2] and BERNICE A. MARCOPULOS[3]

[1]*Department of Neuropsychology, Catholic University of Córdoba, Argentina*
[2]*Department of Psychometric Techniques, National University of Córdoba, Argentina*
[3]*Neuropsychology Lab, Western State Hospital, Staunton, VA, USA*

Fernandez, A. L. & Marcopulos, B. A. (2008). A comparison of normative data for the Trail Making Test from several countries: Equivalence of norms and considerations for interpretation. *Scandinavian Journal of Psychology*, *49*, 239–246.

The Trail Making Test may not be equivalent across cultures, i.e., differences in the scores across different cultures may not reveal real differences in the ability of the subjects on the construct being measured. In order to assess this hypothesis, normative samples from ten different countries were compared. Age decade subgroups across samples were ranked based on mean time taken to complete each part of the task. Large Z scores differences were found between these samples when comparing the first with the second, and the last in the rank. These differences were significant even when age and education were comparable across samples. Following Van de Vijver & Tanzer (1997), several possible sources of bias were identified. Incomparability of samples and administration differences were the most likely factors accounting for differences. Because of the lack of validity studies in the countries considered, no firm conclusions could be obtained regarding construct bias. Although the TMT may be measuring visual scanning, psychomotor speed and mental flexibility, normative data from different countries and cultures are not equivalent which might lead to serious diagnostic errors.

Key words: Cross-cultural neuropsychology, Trail Making Test, test bias, cultural influence, normative data.

Alberto L. Fernandez, Cortex-Neuroterapias, Chacabuco 1296, 1º D 5000, Córdoba, Argentina. Tel: 54-351-4601184; e-mail: neurorehab@onenet.com.ar

INTRODUCTION

The Trail Making Test is one of the most common attention and cognitive flexibility tests used across the world in neuropsychological assessment. Initially created by the US Army (US War Department, 1944), it was later included in the Halstead-Reitan Neuropsychological Battery (Reitan, 1955). The test consists of two parts, one with numbers and the other with numbers and letters. In Part A, the task of the subject is to connect the numbers in ascending order, while in Part B the subject is required to connect alternatively numbers and letters following the ascending order of the numbers and the sequential order of the alphabet letters. The time taken to complete every part is the score to be considered when analyzing the performance. The TMT is considered a test of visual search, attention, mental flexibility and motor function (Spreen & Strauss, 1998). Motor speed has been found to have a strong influence on the performance on this test (Lezak, 1995). Part A, considerably less demanding than Part B, is considered a measure of visual search and mental tracking; whereas Part B is assumed to encompass cognitive abilities, such as flexibility, to shift the course of an ongoing activity and the capacity to deal with more than one stimulus at a time. Mirsky, Anthony, Duncan, Ahearn and Kellam (1991) found that both parts loaded on the "focus-execute" factor of their factorial study. According to them, this factor comprises a visual-perceptual ability as well as the capacity to make skilled manual responses. The TMT has been widely used as a sensitive test for the detection of any kind of brain damage (Lezak, 1995). The TMT is affected by age, education, and IQ, but not by gender (see review by Mitrushina, Boone & D'Elia, 1999).

Attention is a highly biologically determined construct. Research shows that deficits in attention are one of the most prevalent results in almost any disease that compromises brain functioning, including traumatic brain injury, Alzheimer's disease, Parkinson's disease, and epilepsy, among others (Rankin, Adams & Jones, 1996; Sohlberg & Mateer, 1989; Soukup & Adams, 1996; Williamson, Scott & Adams, 1996; Zec, 1993). Since the TMT measures attention, it might be assumed that its measurement is free of cultural influence. Therefore, the use and results of this test should be unaffected whether it is applied in Argentina or Australia. The clinician practicing outside the US, for example, might feel that using North American norms could be legitimate since they would be unaffected by cultural differences.

However, research has shown that cultural variables affect cognitive test performance. A test does not always measure the same construct when it is administered in a different context than where it was developed (Ardila & Moreno, 2001; Greenfield, 1997; Rogoff & Chavajay, 1995). Van de Vijver and Tanzer (1997) identified other sources of bias when a cross-cultural comparison of test results is made, like incomparability of samples. They identified three main sources, namely (1) construct bias; (2) method bias; and (3) item bias.

 Published by Blackwell Publishing Ltd., 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA. ISSN 0036-5564.

In this article, normative data for the TMT across eleven different countries are compared. The aim of this study is to evaluate the equivalence of these norms and to assess the presence or absence of any kind of bias in these normative data for the TMT, according to Van de Vijver and Tanzer's taxonomy.

METHOD

Participants

A total of 11 studies were included in this paper. It was difficult to obtain a comprehensive international sample of normative data because most of these data are not published outside the countries of origin. Also, very few normative studies have been published on the TMT outside North America. The countries and studies reporting their normative data included in this study were: Argentina (Fernández, Marino & Alderete, 2002), Belgium (Lannoo & Vingerhoets, 1997), Canada (Goul & Brown, 1970), China (Lee & Chan, 2000a), Denmark (Nielsen, Knudsen & Daugbjerg, 1989; Nielsen, Lolk & Kragh-Sørensen, 1995), Italy (Giovagnoli, Del Pesce, Mascheroni, Simoncelli, Laiacona & Capitani, 1996), New Zealand (Siegert & Cavana, 1997), Sweden (Bergman, Bergman, Engelbrektson, Holm, Johannesson & Lindberg, 1988), UK (Stewart, Richards, Brayne & Mann, 2001), and USA (Selnes, Jacobson, Machado *et al.*, 1991).

The samples were very heterogeneous in size, the age ranges, and the educational characteristics of the subjects across the articles. Since age is the most important demographic variable in TMT performance (Mitrushina *et al.*, 1999), we attempted to match the age groups (usually grouped by decades) in order to approximate comparable subgroups. Table 1 shows all the studies included. Data regarding the number of subjects, age range and education of the samples are provided. All of the samples included neurologically intact subjects. Although most of the samples were large enough to be considered "normative", the Chinese one had only 35 subjects. Unfortunately, the educational levels of the subjects were not always specified. In many cases there was no detailed explanation about how many subjects of every educational level (i.e. primary, secondary or university) were included. Most of the samples were from countries that could be considered part of the Western culture. Although the definition of "Western culture" is elusive, in this paper we define it as the predominant religious, philosophical, and social values, as well as language family of the societies of West Europe and North America; i.e., Christian religion, rationalistic thinking, Indo-European languages, and individualism over collectivism. In the psychological testing field, "Western culture" has been referred to as "test-wiseness" by Nell (2000). Test-wiseness refers to the knowledge of all the appropriate attitudes in a testing situation: to work fast and accurately, with intense concentration and in silence. In the case of these samples, they also have commonalities regarding an alphabetic characters system (Latin characters are used) and the same numeric system (Arabic). The Chinese sample is an exception. This sample was comprised of university students living in Hong Kong. However, it may also be argued that there is some influence of Western culture because of the long and profound influence of English culture in Hong Kong, which was until recently a British colony.

Some distinctive characteristics of the normative studies included are listed below:

(a) There were two studies from Denmark (Nielsen *et al.*, 1989, 1995); one, which included people aged 20–54 and the other, people aged 64–83. For the people aged 64–83, Part B was not administered.
(b) The Chinese study is not a normative data study but compared the performance on TMT and CTT in a sample of Chinese students.
(c) In the UK study, Part B scores and standard deviations were not reported.
(d) Participants in the study by Stewart *et al.* (2001) were African Caribbeans who emigrated to UK many years ago.
(e) Participants of the Belgian study were Flemish.

Table 1. *Country, sample size, age and education ranges for studies included in the comparisons*

Country	*N*	Age	Education
Argentina[a,b,d]	251	15–70	All educational levels
Belgium[a,b]	200	18–74	7–19 yrs
Canada[a,e]	122	20–72	6–13 yrs
China	35	$M = 20 \pm 0.9$	University students
Denmark[a,b]	156	20–54	Not specified
	134	64–83	80% with 7 or less
Italy[a,b,c,e]	287	20–79	All educational levels
New Zealand[a,e]	127	60–80+	$M = 10+$
Sweden[a,b]	400	20–65	12+ yrs
UK[a,b,c,d]	285	55–75	All educational levels
USA[a]	696	25–54	$M = 16$ yrs

Note: Demographic variables influencing norms: a = Age; b = Education; c = Gender; d = Occupation; e = IQ.

Procedures

Tables 2 and 3 show the means and standard deviations for each age range reported across studies. Countries were ranked according to the performances on Part A and B from the shortest to the longest times (Tables 4 and 5). In order to evaluate the magnitude of the differences between norms, the distances, as expressed in standard deviations, were calculated between the first, and the second and last country in the ranking. Table 6 shows these distances.

Data was inspected to evaluate the influence of different demographic variables. Finally, mean times across age decades were plotted for all the countries (where available) to visualize the influence of age in every sample. (See Figs 1–4.)

RESULTS

Tables 4 and 5 show the results of the ranking of countries for Parts A and B. The US sample by Selnes *et al.* (1991), in the subgroups where included, is always in the first place, i.e. top score, on Part A. On Part B, the US sample is in the first place in two subgroups (30–39 and 40–49), and is in the second place in the 20–29 subgroup. In this latter group the Chinese sample is in the first place. There is only a slight difference between the Chinese and the North American sample for the age group 20–29. The Swedish and New Zealand samples were the top ones in the remaining subgroups. Italian, Argentinean and British samples had the slowest times. The Canadian sample occupies the last place on Part B in the 20–29 subgroup.

Table 2. *TMT A means and standard deviations for each age range across studies*

	Trails A		
	Age	*M* (sec.)	*SD*
Argentina	20–29	38.88	12.55
	30–39	38.68	13.53
	40–49	45.57	13.58
	50–59	54.60	21.35
	60–69	67.88	29.61
Belgium	18–29	27	6
	31–50	33	15
	51–74	39	14
Canada	20–29	36.10	10.00
	30–39	35.50	9.40
	40–49	40.00	13.30
	50–59	45.30	13.60
	60–72	68.90	21.20
China	20–29	24.70	7.80
Denmark	20–29	26.91	10.53
	30–39	29.78	8.65
	40–54	36.64	12.91
	64–69	58.20	20.13
	70–74	67.68	22.79
	75–79	68.59	26.02
	80+	67.00	28.37
Italy	20–29	33.45	13.03
	30–39	38.75	16.40
	40–49	48.90	23.58
	50–59	53.83	26.30
	60–69	67.26	28.69
	70–79	84.60	23.76
New Zealand	60–64	34.76	9.46
	65–69	38.04	14.32
	70–74	45.04	14.03
	75–79	46.78	17.89
	80+	60.21	24.38
Sweden	20–34	28.00	10.70
	35–49	33.20	12.60
	50–65	34.10	12.50
UK	55–64	68.00	Not reported
	65–75	88.00	Not reported
USA	25–34	19.00	5.90
	35–44	20.80	5.50
	45–54	23.10	7.30

Table 3. *TMT B means and standard deviations for each age range across studies*

	Trails B		
	Age	*M* (sec.)	*SD*
Argentina	20–29	72.31	20.67
	30–39	75.27	24.67
	40–49	89.13	31.77
	50–59	104.85	42.73
	60–69	147.38	59.15
Belgium	18–29	60	16
	31–50	73	28
	51–74	76	21
Canada	20–29	85.70	38.70
	30–39	79.60	20.40
	40–49	105.20	42.20
	50–59	103.20	43.30
	60–72	158.80	49.50
China	20–29	44.70	12.00
Denmark	20–29	60.91	19.87
	30–39	63.70	17.73
	40–54	78.79	27.24
Italy	20–29	78.06	33.69
	30–39	86.24	34.39
	40–49	111.77	53.98
	50–59	134.50	80.06
	60–69	164.54	97.41
	70–79	336.80	197.80
New Zealand	60–69	84.57	Not reported
	70–79	125.41	Not reported
	80+	195.93	100.39
Sweden	20–34	64.00	26.00
	35–49	72.50	25.90
	50–65	81.50	36.40
UK	Not reported		
USA	25–34	49.50	17.10
	35–44	52.50	18.60
	45.54	53.90	20.30

Table 4. *TMT A rank order by total time from shortest to longest by age group for each study*

	Age groups						
Order	20–29	30–39	40–49	50–59	60–69	70–79	80+
1	USA	USA	USA	Sweden	New Zealand	New Zealand	New Zealand
2	China	Denmark	Belgium	Belgium	Belgium	Denmark	Denmark
3	Denmark	Belgium	Sweden	Canada	Denmark	Italy	
4	Belgium (Flemish)	Sweden	Denmark	Italy	Italy		
5	Sweden	Canada	Canada	Argentina	Argentina		
6	Italy	Argentina	Argentina	UK	Canada		
7	Canada	Italy	Italy		UK		
8	Argentina						

Table 5. *TMT B rank order by total time from shortest to longest by age group for each study*

	Age groups						
Order	20–29	30–39	40–49	50–59	60–69	70–79	80+
1	China	USA	USA	Belgium	New Zealand	New Zealand	New Zealand
2	USA	Denmark	Sweden	Sweden	Belgium	Italy	
3	Belgium	Sweden	Belgium	Canada	Argentina		
4	Denmark	Belgium	Denmark	Argentina	Canada		
5	Sweden	Argentina	Argentina	Italy	Italy		
6	Argentina	Canada	Canada				
7	Italy	Italy	Italy				
8	Canada						

Table 6. *Differences in standard deviations between the first, and the second and last in the rank*

	PART A		PART B	
Age group	Second	Last	Second	Last
20–29	–1.0	–3.4	–0.4	–3.4
30–39	–1.6	–3.3	–0.6	–1.8
40–49	–1.4	–3.5	–0.9	–2.9
50–59	–0.4	–1.6	–0.3	–2.8
60–69	–0.4	–3.6	0.0	–3.8
70–79	–1.6	–2.8	–3.7	–3.7
80+	–0.3	–0.3		

Note: Differences in standard deviations were calculated between the first country in the rank and the second; and the first country in the rank and the last one.

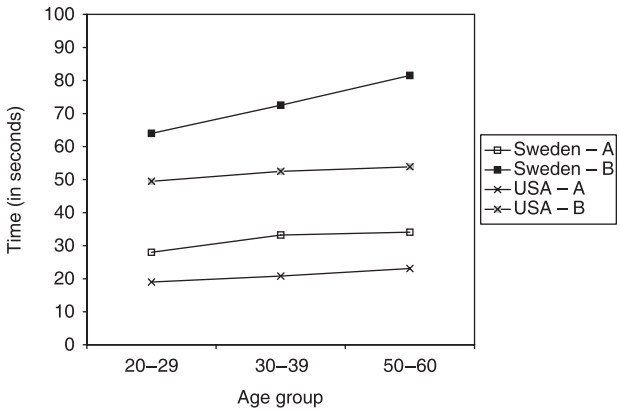

Fig. 1. Performance across decades for the Swedish and American samples.

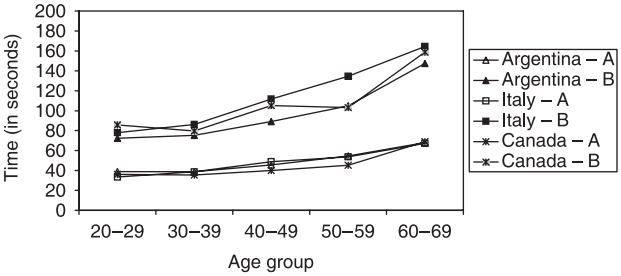

Fig. 2. Means for TMT A and B for Argentina, Italy and Canada by age decade.

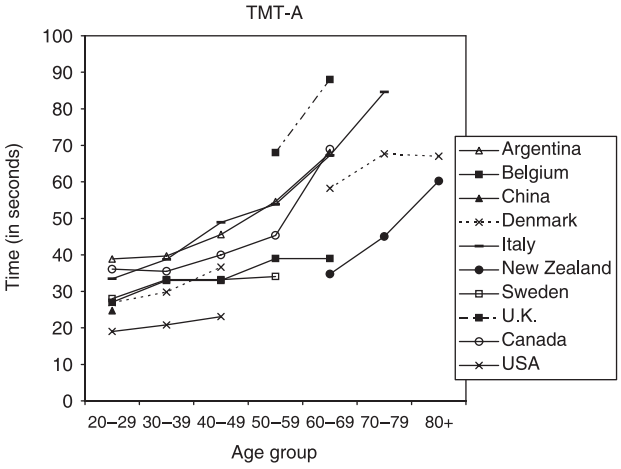

Fig. 3. Performance of all samples on TMT A.

There are large differences between the first and the last in the rank in almost every subgroup. The most striking difference is found in the 60–69 subgroup on Part B, where New Zealanders outperform Italians by 3.8 standard deviations. There are also large differences between the first and the second in the rank, especially on Part A. The largest distance can be found in the 30–39 and 70–79 subgroups where the second country in the rank is 1.6 standard deviations below the first one. Table 6 displays the differences between the first and the second ranked norms by time, and the first and the last ranked norms for each age group. Table 1 includes the demographic variables that were influencing the scores in every study considered. Age and education is the most pervasive influence found on the TMT scores.

Gender influenced scores in the Italian and British studies. Females had slightly longer times than men among the

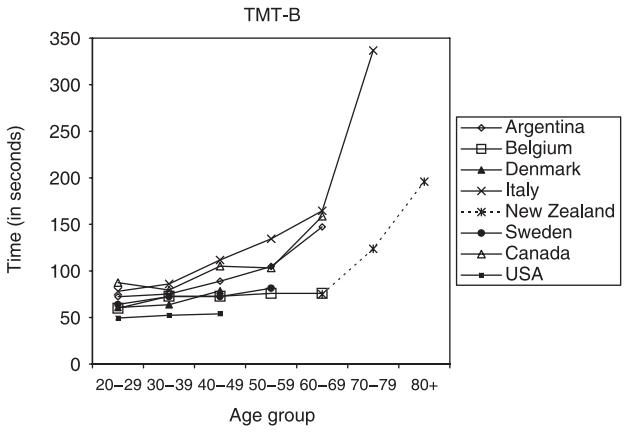

Fig. 4. Performance of all samples on TMT B.

Italians only on Part A. The same results are reported for the British sample. Occupation also had a significant influence on the scores. In the Argentinean sample, even after the adjustment for education, unskilled laborers along with unemployed, retired people and housewives showed the slowest times on both parts. Professionals and students obtained the best scores. In Stewart *et al.* (2001), a relationship between social class and TMT scores was reported. However, they estimated social class from previous occupation. They found that lower social class was associated with lower scores. Intelligence, as measured by different tests, influenced the performance of TMT. The Italian study used the Ravens Colored Progressive Matrices as an estimate of intelligence and reported a large and significant influence of intelligence on the scores of both parts. Siegert and Cavana (1997), in the New Zealand study, found a significant negative correlation of 0.53 for Part A and 0.54 for Part B. In this case, IQ was measured with Vocabulary and Block Design subtests of the Wechsler Memory Scale–Revised.

Two groups were formed with samples that are comparable in size and demographic characteristics, such as education and age. In the first group, the Swedish sample was compared with the North American. In the second group, the Argentinean, the Canadian, and the Italian samples were compared. In the first group, distances vary from 0.8 to 1.4 standard deviations always on Part B, with North Americans obtaining better scores than Swedish. Performance across decades for both samples is plotted in Fig. 1. While the slope is very similar on Part A, on Part B the Swedish sample shows a steeper decline. In the second group, the most striking differences are found between the Argentinean and Italian samples, again on Part B. These distances vary from 0.2 to 0.7 standard deviations. Figure 2 shows the plotting of the mean times by decades on both parts. Inspection of the plot clearly reveals that the subjects of the Canadian sample have a different and inconsistent performance across decades group, especially on Part B. For instance, the 30–39 group has a better performance than the 20–29 group. Likewise, the 50–59 group performed better than the 40–49 group. These abnormalities are not seen in the remaining samples. Figures 3 and 4 show the performance of all the samples on Parts A and B, respectively. On Part A, most of the samples have a similar slope, with an abrupt decline in the sixties. Singularly, the Flemish sample has a very slight decline from twenties to sixties. Remarkable is also the performance by the 80+ group of the Danish sample, which obtained shorter times than the 70–79 group. A similar pattern arises on Part B. Again, the Flemish sample has an unusual decline; i.e., very different from the slope observed in the rest of the samples.

DISCUSSION

The most elementary inspection of these data affirm that, even when considering a basic psychological process such as attention, norms are not interchangeable, even among countries that have a "Western" style of education. As it has been stated above, there are large differences between the norms obtained in different countries. In some cases, these differences are so dramatic that normal subjects could be classified as pathological and vice versa, depending upon the norms used. This is true even in demographically comparable samples, such as the Swedish and the North American.

There are two exceptions to this finding: the Italian and the Argentinean samples, which yielded very similar results, as well as the Swedish and Flemish samples. Nevertheless, this is probably a fortuitous coincidence, i.e., there appear not to be sound theoretical reasons to explain these similarities. Furthermore, distances between some decade groups of these comparable samples reach 0.7 standard deviations. These distances can produce very different results in the evaluation of a given individual. For instance, if an Argentinean were to be assessed with the Italian norms, this person might be considered as having a deficit when his/her performance may be in a normal range when assessed with Argentinean norms.

Therefore, it is justifiable to say that there is a sample bias on TMT between these studies because of the different compositions of education, occupation, and intelligence which can affect TMT score. It would be unwise and untenable to conclude that North Americans have a better attentional capacity than Danish, Swedish or Italians. Soukup *et al.* (1998) found large differences on TMT performance across North American samples. Similarly, Mitrushina *et al.* (1999) list 22 TMT normative studies done in North America in their handbook on norms and show differences among them.

As mentioned earlier, Van de Vijver and Tanzer (1997) identified three possible sources of test bias, including: (1) construct bias; (2) method bias; and (3) item bias. They said that construct bias "occurs if the construct measured is not identical across cultural groups" (p. 264). None of the

studies cited in this paper directly addressed construct validity. Therefore, one cannot make any conclusions on whether there is construct equivalence across the countries represented by the normative studies. Nonetheless, hypotheses regarding the issue of construct equivalence from the TMT can be entertained to guide further research. All of the countries represented by the study are Western cultures that use the same numeric system and alphabet (except by the Chinese group who were bilinguals). Consequently, numeric system and alphabet should not influence the construct being measured in these cases. Besides, the influence of the same demographic variables on the scores and the shape of the slope observed in the graphics also suggest construct equivalence. However, it should be recognized that age and education are very powerful variables largely influencing the scores on any neuropsychological test (Ostrosky-Solis, Ardila, Rosselli, López-Arango & Uriel-Mendoza, 1998). In fact, a difference in the level and prevalence of formal education is often the underlying explanation in test performance differences across cultures (e.g., Levav, Mirsky, French & Bartko, 1998; Nell, 2000). A Trail Making Test that uses colors rather than letters was developed to address some of the difficulties of transporting tests across cultures (D'Elia, Satz, Uchiyama & White, 1994).

Although the data presented in this paper cannot confirm construct equivalence, there are other findings that suggest that TMT might not be measuring the same construct when applied in different cultures where it was created. For example, the study by Lee, Cheung, Chan and Chan (2000) suggested that equivalence of the TMT and a very similar test like the Color Trails Tests (CTT) may be language-specific since the strong correlations between both tests were weaker in a Chinese-English bilingual sample compared to a primarily English speaking sample. Lu and Bigler (2000) tested Chinese students attending an American university and found that the Chinese students took longer to complete Trail B than the American students. These differences disappeared when the Chinese students completed a modified version of Trail B using Chinese characters. Likewise, Lee & Chan (2000b) found in a sample of Chinese individuals that correlations between the TMT and CTT varied across different age and education groups. As a result, they concluded that construct equivalence between TMT and CTT in Chinese may be found under specific age and education parameters. Similarly, Dugbarty, Townes & Mahurin (2000) found statistically significant differences in performance on CTT-2 and TMT Part B, as well as the interference indices for both tests in a sample of 64 English-Turkish bilingual Turkish university students. They interpreted these data as evidence of inequivalence between these parts of the tests.

In all the studies cited above, the English version was applied to bilingual subjects whose other language was not Western, i.e., Indo-European. Other studies have attempted to modify the TMT for non-Western languages as shown in the study by Axelrod, Aharon-Peretz, Tomer and Fisher (2000). In Part B, the Hebrew letters are substituted for the English letters. According to Axelrod *et al.*, "Hebrew letters are usually used in place of numbers in the same manner as are Roman numerals in English (e.g., school grades are labeled by letter [aleph, bet, etc.] rather than by ordinal number)" (p. 187). Then, while the 20 first items are rather simple for Israelis, after number 10, the task increases its difficulty dramatically since numbers greater than 10 are a combination of letters instead of a single letter. Therefore, in this case it is very likely that cognitive flexibility is not what is being measured before number 10 in this part of the TMT. Thus, although there is no evidence of construct inequivalence across the studies included in this research, data from other studies warn about the possibility of construct bias. It remains a question whether this bias is only present when the test is administered to non-Western individuals.

Sample bias is a subtype of method bias that occurs when samples are not comparable. The samples cited in this paper differed on many important demographic factors, such as age and education, and prevented a true comparison of the TMT. Thus, one explanation for the enormous differences between the normative data presented here is incomparability of samples. For instance, 80% of the aged Danish subjects had 7 or less years of education, while the New Zealanders had a mean of more than 10 years of education. Educational composition of the samples can have a remarkable influence in the results. Many subjects in these samples have only primary school. Nell (2000) underscored the powerful influence of the schooling process in the development of cognitive process and its consequences in neuropsychological testing. He suggested that constructs can vary according to the educational level of the subject. On Part B, the test attempts to measure stimulus resistance, i.e., the ability to inhibit the automatic tendency to move to the next stimulus in the same entrenched series. For instance, A suggests moving to B, and B suggests moving to C, and so on. Similarly, 1 suggests moving to 2, 2 to 3, and so forth. Since the recitation of the alphabet is unpracticed in semi-literate subjects, they may have little conflict in switching from numbers to letters. Thus, the stimulus resistance effect is diminished. Data obtained by Fernández *et al.* (2002) tend to confirm this modification of the construct in lower educational levels. They found that the correlations among three "executive functioning tests" [Porteus Maze Test, Controlled Oral Word Association Test, and TMT] were weaker or even absent in the groups of subjects with less than eight years of education. Ostrosky-Solis *et al.* (1998) found that a few years of education had a major impact on the performance on a neuropsychological battery. Subjects with 3 or 4 years of education had a significantly higher performance than illiterate subjects, especially in verbal tasks. Interestingly, the TMT has a verbal component.

Other demographic characteristics were different across samples to cause bias. For instance, the subjects in the North American study were all men. However, there are some of

these samples that are comparable and yet they had very different mean scores, as shown above. Thus, other sources of bias are probably operating in this case. For instance, there are probable differences in health and social variables that may account for some differences. Also, neurological exclusion data and the definition of "normal" likely differ among samples.

Van de Vijver and Tanzer also identified the administration bias, which is caused by the particular form of administration. The most widely used method of administration of the TMT is to stop the subject as he/she makes mistakes and ask him/her to correct them (Spreen & Strauss, 1998). The scoring system only takes time into account. Other systems include allowing the subject to perform with errors, but giving a low score for uncorrected errors. Unfortunately, in many of the studies included in this article, there was no indication about what administration system was used. This could be another important source of bias. In addition, it has been reported that differences in response times and correction styles of the administrators can significantly affect reliability (Snow, 1987; cited in Lezak, 1995). Nielsen *et al.* (1989) provided the administration system they used. They reported that "the stopwatch was started when the patient initiated the first stroke . . ." (p. 39). This is not the same procedure used in the Argentinean research where the stopwatch, according to Spreen and Strauss (1998), was started once the administrator placed the sheet in front of the subject after the trial. This can result in potentially significant time differences. Many subjects take some time looking carefully at the page before they initiate a stroke. Then, once they start the line, they can perform faster since they had a preview of the numbers and letters arrangement. This might explain, in part, the enormous differences between the Argentinean and Danish data.

Within method bias, Van de Vijver and Tanzer also considered stimulus familiarity as another source of bias. Lee *et al.* (2000) investigated if the particular array of the stimuli in the TMT had an impact on speed of performance between different cultural groups. In the TMT, completion of the task mainly involves moving back and forth in an up-down direction. Therefore, they hypothesized that Chinese-English bilinguals would be faster than Americans in completing the TMT since previous research had demonstrated that English speakers were slower to read letters presented in vertical columns than if they were arranged in horizontal rows. This effect had not been seen in Chinese speakers. This is probably related to the vertical direction of the Chinese reading system. Their results were contrary to this hypothesis since their English speakers were as fast as their Chinese-English bilingual speakers. They concluded that TMT is a "fair" test with respect to language background.

In summary, there is a remarkable variability in TMT performance between the studies of the countries analyzed in this research. This reinforces the obvious conclusion that a clinician cannot use normative data on a test interchangeably. The clinician should ensure that he/she is using the same administration procedure used in the normatization study and that the subject under evaluation fits the normative sample in some basic demographic characteristics, such as age, education, and gender. In our review of TMT studies across cultures, the most likely source of "bias" (or variability in the normative values across countries) is method bias. Within method bias, incomparability of samples and administration differences may be the main reason the norms are not comparable and do not reflect real differences in the underlying construct. Although not directly measured, construct bias might also be present at some degree and should be studied further.

There are a number of limitations in this study which should be mentioned. Although the participants across studies were grouped by age, there may have been important differences between the samples within the age group. For instance, there may be health differences due to variable health care and social services. The educational system may differ across country samples. Neurological screening criteria for each study may have been different.

This research has shown that there is wide variability in TMT norms across several "Western" culture countries. Whether there is also a cultural bias in construct validity was not amenable to a deep analysis. Analyzing this was not possible with these data because of method bias and constitutes a limitation of the present study. It is difficult to separate the impact of culture from differences in test administration. In current cross-cultural neuropsychology, it is important to assess if the constructs are constant across cultures; mainly, if the tests used to measure that construct in the culture where they were created are measuring the same construct in different cultures. Future research should be carried out with comparable samples across countries, particularly comparing Western and non-Western countries in order to assess this issue. Reliable and valid norms need to be developed across cultures to move the science of clinical neuropsychology forward.

The authors would like to thank Joyce Manley, MS, for her assistance in preparing tables and figures for this manuscript.

REFERENCES

Ardila, A. & Moreno, S. (2001). Neuropsychological test performance in Arauco Indians: An exploratory study. *Journal of the International Neuropsychological Society*, *7*, 510–515.

Axelrod, B., Aharon-Peretz, J., Tomer, R. & Fisher, T. (2000). Creating interpretations guidelines for the Hebrew Trail Making Test. *Applied Neuropsychology*, *7*(3), 186–188.

Bergman, H., Bergman, I., Engelbrektson, K., Holm, L., Johannesson, K. & Lindberg, S. (1988). *Psykologhandbok vid alkoholkliniken*. (Manual for psychologists.) Karolinska sjukhuset, 4:e upplagan.

Castro-Caldas, A., Petersson, K., Reis, A., Stone-Elander, S. & Ingvar, M. (1998). The illiterate brain. Learning to read and write during childhood influences the functional organization of the adult brain. *Brain*, *121*, 1053–1063.

D'Elia, L., Satz, P., Uchiyama, C. & White, T. (1994). *Color Trails Test: Professional manual*. Odessa, FL: PAR.

Dellatolas, G., Deloche, G., Basso, A. & Claros-Salinas, D. (2001). Assessment of calculation and number processing using the EC301 battery: Cross-cultural normative data and application to left- and right-brain damaged patients. *Journal of the International Neuropsychological Society*, *7*(7), 840–859.

Dugbartey, A., Townes, B. & Mahurin, R. (2000). Equivalence of the Color Trails Test and Trail Making Test in nonnative English-speakers. *Archives of Clinical Neuropsychology*, *15*(5), 425–431.

Fernández, A., Marino, J. & Alderete, A. (2002). Estandarización y validez conceptual del Test del Trazo en una muestra de adultos argentinos [Normative data and conceptual validity of the Trail Making Test in a sample of Argentinean adults]. *Revista Neurológica Argentina*, *27*, 83–88.

Giovagnoli, A., Del Pesce, M., Mascheroni, S., Simoncelli, M., Laiacona, M. & Capitani, E. (1996). Trail Making Test: normative values from 287 normal adult controls. *Italian Journal of Neurological Sciences*, *17*, 305–309.

Goul, W. & Brown, M. (1970). Effects of age and intelligence on Trail Making Test performance and validity. *Perceptual and Motor Skills*, *30*, 319–326.

Greenfield, P. M. (1997). You can't take it with you. Why ability assessments don't cross cultures. *American Psychologist*, *52*, 1115–1124.

Lannoo, E. & Vingerhoets, G. (1997). Flemish normative data on common neuropsychological tests: Influence of age, education and gender. *Psychologica Belgica*, *37*(3), 141–155.

Lee, T. & Chan, C. (2000a). Comparison of the Trail Making Test and Color Trails Tests in a Chinese context: a preliminary report. *Perceptual and Motor Skills*, *90*(1), 187–190.

Lee, T. & Chan, C. (2000b). Are Trail Making and Color Trails Tests of equivalent constructs? *Journal of Clinical and Experimental Neuropsychology*, *22*(4), 529–534.

Lee, T., Cheung, C., Chan, J. & Chan, C. (2000). Trail Making across languages. *Journal of Clinical and Experimental Neuropsychology*, *22*(6), 772–778.

Levav, M., Mirsky, A. F., French, L. M. & Bartko, J. J. (1998). Multinational neuropsychological testing: Performance of children and adults. *Journal of Clinical and Experimental Neuropsychology*, *20*, 658–672.

Lezak, M. (1995). *Neuropsychological assessment* (3rd edn). New York: Oxford University Press.

Lu, L. & Bigler, E. D. (2000). Performance on original and on a Chinese version of Trail Making Part B: A normative bilingual sample. *Applied Neuropsychology*, *7*, 243–246.

Mirsky, A., Anthony, B., Duncan, C., Ahearn, M. & Kellam, S. (1991). Analysis of the elements of attention: A neuropsychological approach. *Neuropsychology Review*, *2*, 109–145.

Mitrushina, M. N., Boone, K. B. & D'Elia, L. F. (1999). *Handbook of normative data for neuropsychological assessment*. New York: Oxford University Press.

Nell, V. (2000). *Cross-cultural neuropsychological assessment. Theory and practice*. Hillsdale, NJ: Lawrence Erlbaum.

Nielsen, H., Knudsen, L. & Daugbjerg, O. (1989). Normative data for eight neuropsychological tests based on a Danish sample. *Scandinavian Journal of Psychology*, *30*, 37–45.

Nielsen, H., Lolk, A. & Kragh-Sørensen, P. (1995). Normative data for eight neuropsychological tests, gathered from a random sample of Danes aged 64 to 83 years. *Nordisk Psykologi*, *47*(4), 241–255.

Ostrosky-Solis, F., Ardila, A., Rosselli, M., López-Argango, G. & Uriel-Mendoza, V. (1998). Neuropsychological test performance in illiterate subjects. *Archives of Clinical Neuropsychology*, *13*(7), 645–660.

Rankin, E. J., Adams, R. L. & Jones, H. E. (1996). Epilepsy and nonepileptic attack disorder. In R. L. Adams, O. A. Parsons, J. L. Culberston & S. J. Nixon (Eds.), *Neuropsychology for clinical practice*. Washington DC: American Psychological Association.

Reitan, R. M. (1955). The relationship of the Trail Making Test to organic brain damage. *Journal of Consulting Psychology*, *19*, 393–394.

Rogoff, B. & Chavajay, P. (1995). What's become of the research on the cultural bias of cognitive development? *American Psychologist*, *50*, 859–877.

Selnes, O., Jacobson, L., Machado, A., Becker, J., Wesch, J., Miller, E., Visscher, B. & McArthur, J. (1991). Normative data for a brief neuropsychological screening battery. *Perceptual and Motor Skills*, *73*, 539–550.

Siegert, R. & Cavana, C. (1997). Norms for older New Zealanders on the Trail Making Test. *New Zealand Journal of Psychology*, *26*(2), 25–31.

Sohlberg, M. M. & Mateer, C. A. (1989). *Introduction to cognitive rehabilitation*. New York, EE. UU: The Guilford Press.

Soukup, V. M. & Adams, R. L. (1996). Parkinson's disease. In R. L. Adams, O. A. Parsons, J. L. Culberston & S. J. Nixon (Eds.), *Neuropsychology for clinical practice*. Washington DC: American Psychological Association.

Soukup, V., Ingram, F., Grady, J. & Schiess, M. (1998). Trail Making Test: issues in normative data selection. *Applied Neuropsychology*, *5*(2), 65–73.

Spreen, O. & Strauss, E. (1998). *A compendium of neuropsychological tests: Administration, norms and commentary* (2nd edn). New York: Oxford University Press.

Stewart, R., Richards, M., Brayne, C. & Mann, A. (2001). Cognitive function in UK community-dwelling African Caribbean elders: Normative data for a test battery. *International Journal of Geriatric Psychiatry*, *16*, 518–527.

US War Department, Adjutant General's Office (1944). *Army Individual Test Battery: Manual of directions and scoring*. Washington, DC: Author.

Van de Vijver, F. & Tanzer, N. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *47*, 263–279.

Williamson, D. J. G., Scott, J. G., & Adams, R. L. (1996). Traumatic brain injury. In R. L. Adams, O. A. Parsons, J. L. Culberston & S. J. Nixon (Eds.), *Neuropsychology for clinical practice*. Washington DC: American Psychological Association.

Zec, R. F. (1993). Neuropsychological functioning in Alzheimer's disease. In R. W. Parks, R. F. Zec & R. S. Wilson (Eds.), *Neuropsychology of Alzheimer's Disease and other dementias*. (pp. 3–80). New York: Oxford University Press.

Received 14 August 2007, accepted 7 September 2007