Archives of Assessment Psychology, Vol. 6, No. 1, (33-48)

Sociocultural and Language Differences on the Trail Making Test

Ondrej Bezdicek, Ph.D., Ladislav Moták, Ph.D., David J. Schretlen, Ph.D. , Marek Preiss, Ph.D., Bradley N. Axelrod, Ph.D., Tomáš Nikolai, Ph.D., Javier Peña, Ph.D., Natalia Ojeda, Ph.D., and Evžen Růžička, D.Sc.

Abstract

The aim of this study was to test whether a Czech version of the Trail Making Test (TMT) yields normative data scores that approximate those based on English (North American) and Spanish normative samples. In all samples, TMT Parts A and B were administered according to the guidelines. Completion times for healthy adults from the Czech Republic (n = 422, Bezdicek et al., 2012), the United States (n = 325, Schretlen et al., 2008) and Spain (n = 348, Ojeda, Peña, unpublished data), stratified for age and education, were compared. Czechs took less time to complete TMT-B than English and Spaniards in 70–74 and older age groups (all ps < .01) even after covarying for education, although the amount of variance explained by language (i.e. Czech/Spanish/English) was small (0.3%; p < .001). In addition, TMT-A performance was significantly faster in Czech and English (North American) populations than in Spaniards. In conclusion, this demonstrates a lack of sociocultural equivalence on TMT-A and B, suggesting the need for adjustment of available TMT norms for use in different cultures.

Introduction

The Trail Making Test (TMT) was developed in 1938 by Partington and Leiter as a "distributed attention" test and published as part of the Army Individual Test Battery (Partington & Leiter, 1949; U.S. War Department, 1944). It is currently the most frequently used attention test in clinical neuropsychology (Rabin, Barr, & Burton, 2005). It also requires sequencing, visual scanning, processing speed, and mental flexibility for successful performance (Strauss et al., 2006, Tombaugh, 2004).

Evidence suggests that many characteristics of healthy respondents contribute to performance on the TMT, and that some of these include cultural differences (Fernández & Marcopulos, 2008; Loewenstein, Argüelles, Argüelles, & Linn-Fuentes, 1994; Soukup, Ingram, Grady, & Schiess, 1998). However, relatively few studies have assessed the effects of either sociocultural differences or language differences on normative data (Ojeda, Aretouli, Peña, & Schretlen, 2014). As a result, the comparability of TMT norms derived from different countries remains unclear. In this study the term "language" was defined as a system "comprised of symbols (i.e., vocabulary/words), which are socially agreed upon by members of a given community/culture (e.g., American English, French, Spanish, etc.), ..." (Dietz, 2011, p. 1425). The aim of this study was to compare normative data for the TMT from three countries in which the examinees spoke three different languages. Two of the normative databases compared in this study have been published elsewhere and one is thus far unpublished. The English (North American) sample is based on a study described by Schretlen et al. (2008). Normative data derived from this study have been published elsewhere (Schretlen et al., 2010). The Czech normative sample is based on a study described by Bezdicek et al. (2012). The Spanish sample (Ojeda, Peña) is based on unpublished data.

One question is whether English (language) normative data from North America can routinely be adopted by clinicians in different countries, as is often done in the Czech Republic for various tests, including the TMT (Urbánek, 2010), and test the hypothesis that the generalizability of norms outside of the normative sample is questionable (Soukup, Ingram, Grady, & Schiess, 1998). A second goal of our study was to confirm the validity of the Czech normative data for future research in which slight differences in TMT standard scores might play a significant role in differential reasoning, e.g., cutoff scores in the context of mild cognitive impairment (refer to Albert et al., 2011; Arsenault-Lapierre, Whitehead, Belleville, Massoud, Bergman, & Chertkow, 2011; Litvan et al., 2012).

We are aware that multi-cultural comparisons are complicated by many potential methodological differences (Lageman, 2011). For instance, in respect to the TMT, it has been shown that at least some of the compared normative studies manifested incomparability of samples, differences in administration, and also substantially different numeric and alphabetical systems across different languages (Ardila & Moreno, 2001; Axelrod, Aharon-Peretz, Tomer, & Fisher, 2000; Fernández & Marcopulos, 2008; Greenfield, 1997). The present study avoids some of these pitfalls by using relatively similar, if not identical, alpha-numeric systems, identical test formats, and samples of healthy adults. Such an approach is well-suited to disentangle the possible influence of language bias or sociocultural bias on the TMT performance as a subset of "method bias" and "construct bias" defined by Van de Vijver and Tanzer (2004, p. 124) as: "Differential appropriateness of the behaviors associated with the construct (e.g., skills do not belong to the repertoire of the cultural groups)."

Method

Participants. The Czech normative data for the TMT parts A and B consisted of 422 healthy Caucasian subjects aged 20–85 years (for demographic characteristics see Table 1), recruited through local advertisements from Czech communities (for a detailed description of the sample see Bezdicek et al., 2012). The English (North-American) normative data consisted of 325 community-dwelling individuals aged 18–92 years, recruited from the Baltimore, Maryland and Hartford, Connecticut metropolitan areas for a study of normal aging (Schretlen, Munro, Anthony, & Pearlson, 2003). All were recruited as a part of the Johns Hopkins Aging, Brain Imaging, and Cognition (ABC) study (for demographic characteristics see Table 1). In short, the final sample was broadly representative of normal community-dwelling adults (for a detailed description of the sample see Schretlen et al., 2008; Schretlen et al., 2010). Halsted-Reitan normative data was not used, because the normative values were collected several years ago (Reitan & Wolfson, 1993) and may suffer from a strong Flynn effect (Dickinson & Hiscock, 2011). The Spanish normative data consisted of 348 healthy Basque subjects (i.e., Caucasian, because here we report race/ethnic background) aged 15–82 years, recruited through advertisements in local newspapers from communities in Spain (unpublished data). All participants from the three above-mentioned studies were required to have no history of brain damage, psychiatric illness, chronic drug or alcohol abuse, or any medical illness that could affect neurocognitive function. Objective cognitive abilities were within normal limits (i.e., not more than 1,5 SD below age- and education-adjusted normative values in neuropsychological testing), and there were no reports of subjective memory complaints in the healthy control cohort (Schretlen et al., 2010, Bezdicek et al., 2012). All participants were familiar with the respective test language. Altogether, 325 English (North American), 348 Spanish, and 422 Czech participants (*N* = 1095) contributed data to the analyses reported here (Table 1).

Table 1

Basic Demographic Characteristics and TMT Performance in All Three Samples

	EN (*n* = 325)				S (*n* = 348)				CZ (*n* = 422)				Total (*n* = 1095)			
Variables	*M*	*SD*	*min*	*max*	*M*	*SD*	*min*	*max*	*M*	*SD*	*min*	*max*	*M*	*SD*	*min*	*max*
Age	54.72	18.86	18	92	36.32	16.75	15	82	47.69	17.28	20	85	46.16	19.05	15	92
Education	14.22	3.03	3	20	12.45	3.58	0	18	14.19	3.31	6	28	13.65	3.42	0	28
Gender (% female)	56.60	-	-	-	51.70	-	-	-	61.80	-	-	-	57.10	-	-	-
Handedness (% R, L, A)	89.9; 9.0; 1.2	-	-	-	96.6; 3.4;-	-	-	-	89.2; 5.9; 4.9	-	-	-	91.5; 6.2; 2.3	-	-	-
TMT-A	34.91	17.06	13	114	38.43	25.85	13	256	32.85	12.81	12	118	35.24	19.15	12	256
TMT-B	95.34	69.87	31	480	81.63	57.91	21	435	81.59	44.39	25	442	85.68	57.51	21	480
TMT B-A	60.42	60.35	-13	433	43.20	37.48	-11	242	48.74	38.23	1	379	50.45	46.19	-13	433
TMT B/A	2.75	1.28	0.75	10.22	2.19	0.72	0.85	4.80	2.54	0.96	1.02	7.89	2.49	1.03	0.75	10.22

Note. Groups: EN, English normative sample; S, Spanish normative sample; CZ, Czech normative sample; Handedness: R, right-handed, L, left-handed, A, ambidextrous; TMT, Trail Making Test; M, mean; SD, standard deviation; min/max, minimum and maximum value; age and education in years; TMT-A and TMT-B completion times in seconds, derived indices: TMT (B - A) difference score (subtraction of TMT-A from TMT-B; Heaton, Nelson, Thompson, Burks, & Franklin, 1985); the TMT (B/A) ratio score (TMT-B completion time divided by TMT-A completion time; Golden, Osmon, Moses, & Berg, 1981).

Materials. All participants were administered parts A and B of the TMT as a paper and pencil test. In part A of the task, subjects were required to make pencil lines in proper order between 25 randomly dispersed and encircled Arabic numerals arranged on a page (21 × 29 cm). The second part of the task (Part B), consisted of 25 encircled Arabic numerals and Latin letters in alternating order from A to K (L is excluded in the Czech version). In Czech orthography there is a "CH" digraph; a pair of characters used to write one phoneme (a distinct consonant /x/ in the International Phonetic Alphabet, e.g., "prach" [prax], "dust") that does not correspond to the normal values of the two characters combined. This digraph is considered an individual letter, has its own place in the alphabet (after "H" and before "I"), and cannot be separated into constituent graphemes. The Czech and Spanish versions of the TMT-A and B are based on the original format of the English version (Preiss & Preiss, 2006; Reitan & Wolfson, 1993) and both parts are of the same length. TMT-B is different only with regard to the letter (CH). TMT-A and B time to completion were the basic variables of comparison in all studies.

Procedure. All participants completed TMT as part of test batteries described elsewhere (Bezdicek et al., 2012; Ojeda, Peña, unpublished data; Schretlen et al., 2010). In summary, all participants in the three studies were instructed in the aims and procedures of the study, and provided signed, informed consent. The studies were approved by the relevant local medical ethics committee. The TMT was administered in a standardized manner as part of the neuropsychological evaluation and all participants were deemed to have provided sufficient motivation and effort. The administration of the TMT followed the procedures outlined in Strauss, Sherman, and Spreen (2006, p. 656). The total score for TMT-A and TMT-B were measured as the total time in seconds required to complete both tasks (e.g., summation of total completion time of A and then separately of B). If subjects made an error(s), the examiner immediately called it to their attention, and then they had to proceed from the point at which the mistake occurred. Time did not stop during errors and correction of errors (Preiss et al., 2007; Strauss, Sherman, and Spreen, 2006), and errors therefore were reflected in the total completion time rather than as a separate index (Reitan & Wolfson, 1993). We computed mean total completion time (in seconds) for the TMT-A and TMT-B, and we derived the difference (TMT-B - A) and ratio (TMT-B/A) scores which are computed for the measurement of executive control and set shifting in a manner that is independent of psychomotor speed and visual scanning (Bezdicek et al., 2012; Hester, Kinsella, Ong, & McGregor, 2005; Lamberty et al., 1994; Schretlen et al., 2003). All the assessments were done by trained psychometrists in each of the laboratories included in the present study.

Statistical analyses. The threshold significance of p was set at $\leq .05$. The effect size is indicated in terms of partial eta squared (η^2). To interpret the strength of η^2 we followed the guidelines proposed by Tabachnick and Fidell (2007, p. 55): η^2 (% of variance explained .01 or 1% small; .06 or 6% medium; .138 or 13.8% large). All statistical analyses were performed using IBM SPSS Statistics software, version 19.0 (SPSS Inc., Chicago, IL, USA). Given the positive skewness of TMT data (i.e., elongated tail of the distribution due to an excess of participants who took very long to complete the task), all scores were first transformed using logarithmic transformation (log 10). Analyses were then performed using these logarithmic scores. However, for ease of understanding, only raw scores are shown in tables. Positive skewness of TMT data is a common finding (e.g., Tombaugh, 2004).

Results

Correlation analysis revealed that performance on TMT A and TMT B respectively correlated with age (rs = .46 and .48; ps < .001), education (rs = -.26 and -.32; ps < .001), and race (Caucasian/African-American/Other; Spearman's *rhos* = .07 and .17; *ps* < .02). Age thus accounted for 21% and 23%, education for 7% and 10%, and race for 0.5% and 3% of the variance in TMT A and B performance, respectively. Despite being statistically significant, the contribution of race was weak. Thus, only age and education were retained as covariates in subsequent analyses.

Normative data should not differ between comparable alphabetical and numeric systems. Table 2 (see appendix) represents the average performance for each language following different age categories in completion times as well as TMT derived indices (TMT-B-A and B/A). Figure 1 shows that while TMT-A completion times remained almost identical for the English (North American) and Czech samples increasing progressively with age, this age-related increase in TMT-A completion times was steeper in the Spanish sample from 45 years of age onwards. In fact, beyond the main effects of language [$F(2, 1064) = 92.67$, $p < .001$, $\eta^2 = .15$] and age [$F(9, 1064) = 57.30$, $p < .001$, $\eta^2 = .33$], a 3 (language) x 10 (age) analysis of covariance (ANCOVA) performed on TMT-A as a dependent variable and education as a covariate yielded an interaction between both factors [$F(18, 1064) = 7.77$, $p < .001$, $\eta^2 = .17$].

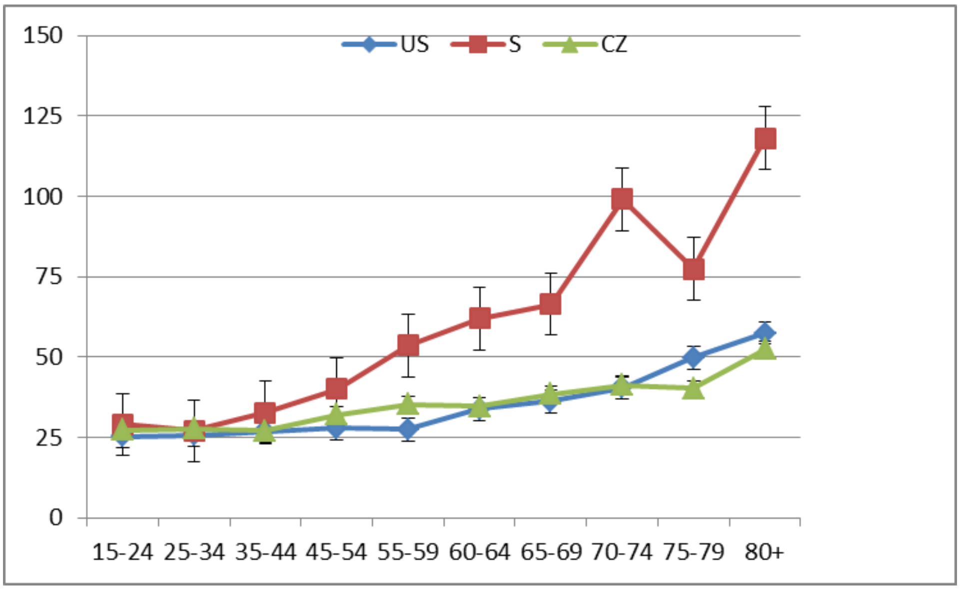

Figure 1. Performance on TMT-A based on "estimated means" (i.e. means based on ANCOVA approach. including education as covariate)
Note. Groups: EN, English normative sample; S, Spanish normative sample; CZ, Czech normative sample; whiskers are standard errors of the mean.

A similar result in TMT-B performance is shown in Figure 2. From 45 years of age onwards, completion times of the Spanish sample increased more rapidly with age than in the English (EN according to language system used, geographically North American) and Czech samples (CZ). Figure 2 also shows that from 70 years of age onwards, completion times of the US sample are at least descriptively longer than completion times of the Czech sample. It can be statistically proven that this is the case in the 70–74 years group ($p < .01$). Unfortunately, other inferential analyses cannot be performed given lower sample sizes per age-case in the remaining age groups (75–79 and 80+ years). An ANCOVA yielded main effects of language [$F(2, 1064) = 21.73$, $p < .001$, $\eta^2 = .04$], age [$F(9, 1064) = 41.17$, $p < .001$, $\eta^2 = .26$], and an interaction between both factors [$F(18, 1064) = 5.67$, $p < .001$, $\eta^2 = .09$].

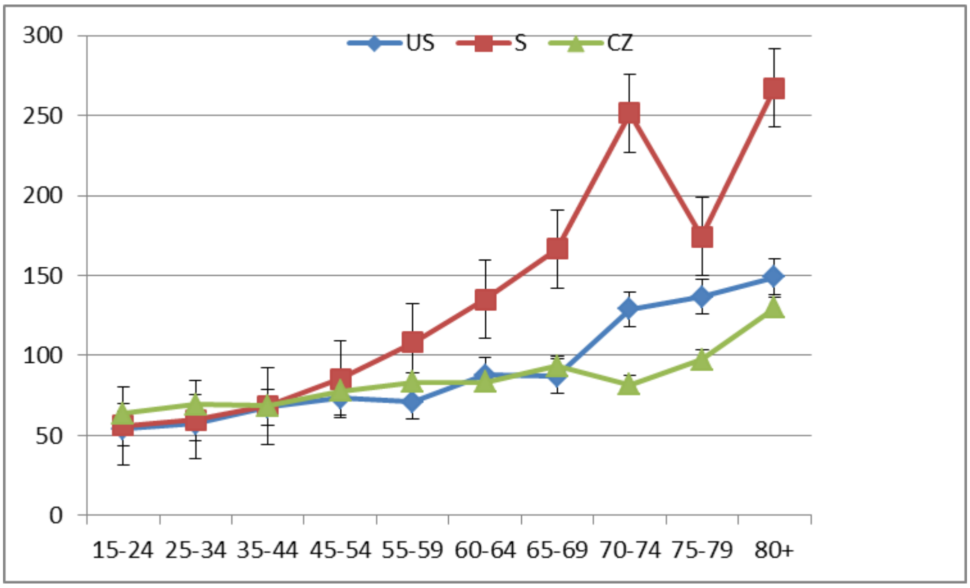

Figure 2. Performance on TMT-B based on "estimated means" (i.e. means based on ANCOVA approach, including education as covariate)
Note. Groups: EN, English normative sample; S, Spanish normative sample; CZ, Czech normative sample; whiskers are standard errors of the mean.

In order to quantify the contribution of different languages to TMT performance, regression analysis on TMT completion times was performed, including all demographic data available. Following the previous example (Testa et al., 2009), the first block of independent variables contained not only handedness (1=Right, 2=Left), race trichotomy (1=Caucasian, 2=African-American, 3=Other), gender (1=Male, 2=Female), age, and education, but also age x education interaction, age-squared and education-squared, in order to account for possible interaction and non-linear effects. Language (1=(American) English, 2=Spanish, 3=Czech) was then added as the second block of independent categorical variables.

In analysis of the TMT-A, the overall model accounted for 32% of the variance [*adjusted* $R^2 = .316$, $F(9, 1054) = 55.48$, $p < .001$]. However, only the first block of variables (handedness, race, gender, age and education) contributed significantly to the model, accounting for nearly all the explained variance ($p < .001$). Adding the term for language did not improve the model ($p = .11$). In contrast, adding a term for language did improve the model for TMT-B performance. Here, the overall model accounted for 40% of the variance [*adjusted* $R^2 = .395$, $F(9, 1054) = 77.98$, $p < .001$], with both the first block (39.7%, $p < .001$) and the second block ((language (English/Spanish/Czech); 0.3% $p < .001$)) making significant contributions. In the case of TMT-B, language differences were found to make a small but reliable contribution to performance (for regression-based norms are listed in Table 3.

Table 3

Regression-based norms including only reliably contributing factors

TMT-A = 52.410 + 0.056*Race – 0.636*Education + 1.016*Age^2 + 0.631*Education^2 - 0.390*Age*Education (SEE = ± 8.748)

TMT-B = 109.108 + 0.188*Race – 0.579*Education + 0.766*Age^2 + 0.512*Education^2 - 0.403*Age*Education + 0.054*Language (SEE = ± 24.942)

Note. SEE = standard error of estimate.

To corroborate evidence of sample differences in TMT-A and B from the previous analysis and to secure that the findings are not unreliable as a result of insufficient sample size in older age groups in accordance with standards set by Mitrushina and colleagues (2005, p. 70) we provide additional samples that combine larger age groups: a “General Adult” sample using individuals ages 25–54, “Older Adults” ages 55–69, and “Elderly” ages 70–80+. In addition to the reliable main effects, all Fs(2, 899) ≥ 10.03, ps < .001, η^2s ≥ .02 (except for age on B/A scores, not significant), a series of complementary 3 (sample) x 3 (age category) ANCOVAs performed on all four TMT scores (A, B, B-A, B/A) with education as a covariate detected reliable interactions, all four Fs(4, 899) ≥ 4.03, all ps < .01, η^2s > .01.

As for the TMT-A, scores consistently increased with participants’ age across all samples (Table 4), such that the lowest scores were achieved by participants aged 25–54 years, somewhat greater scores by participants aged 55–69 years, and the greatest scores by participants aged 70+ years (all ps ≤ .017; Sidak’s multiple comparisons adjustment). Consistently across all age groups, the TMT-A scores did not differ between EN and CZ samples (all ps > .20). Yet, whatever the age group, S participants had reliably greater scores than both EN and CZ participants (all ps < .01).

As for the TMT-B, scores consistently increased with age in both EN and S samples (all ps ≤ .016; Table 4). However, while there was an observable increase in TMT-B from 25–54 to 55–69 age groups in the CZ sample (p < .01), there was no difference in TMT B between 55–69 and 70+ age groups (p = .36). As shown in Figure 3, here were no reliable sample-related differences in the 25–54 age group (all ps > .20). Yet, while EN and CZ participants did not differ in the 55–69 age group (p = .997), they both outperformed (i.e., had lower TMT-B scores) the S participants (all ps < .01). Finally, the most striking differences were found in the highest age level, such that CZ participants outperformed EN participants (p < .001) and both CZ and EN participants outperformed those from Spain (all ps < .05).

Table 4
Reanalysis with another samples: mean marginal estimates (education = 13.77 years) and standard deviations of TMT, part A and B scores following different Sample x Age categories

Parts	Samples	Age groups					
TMT-A		25–54 years		55–69 years		70+ years	
		M	*SD*	*M*	*SD*	*M*	*SD*
	EN (*ns* = 140 / 81 / 88)	27.71	13.82	33.48	12.04	50.53	12.28
	S (*ns* = 156 / 42 / 15)	33.69	14.20	58.34	10.28	100.48‡	7.97
	CZ (*ns* = 222 / 112 / 53)	29.68	15.52	36.45	13.05	43.88	10.82
TMT-B		25–54 years		55–69 years		70+ years	
		M	*SD*	*M*	*SD*	*M*	*SD*
	EN (*ns* = 140 / 81 / 88)	72.56	24.46	89.13	21.32	151.89	21.74
	S (*ns* = 156 / 42 / 15)	71.42	25.14	118.99	18.20	217.88‡	14.11
	CZ (*ns* = 222 / 112 / 53)	78.39	27.47	90.25	23.10	98.90	19.16

Note. Groups: EN, English normative sample; S, Spanish normative sample; CZ, Czech normative sample. ‡ = significant caution to the reader due to small sample size in the Spanish sample (*N* = 15 subjects).

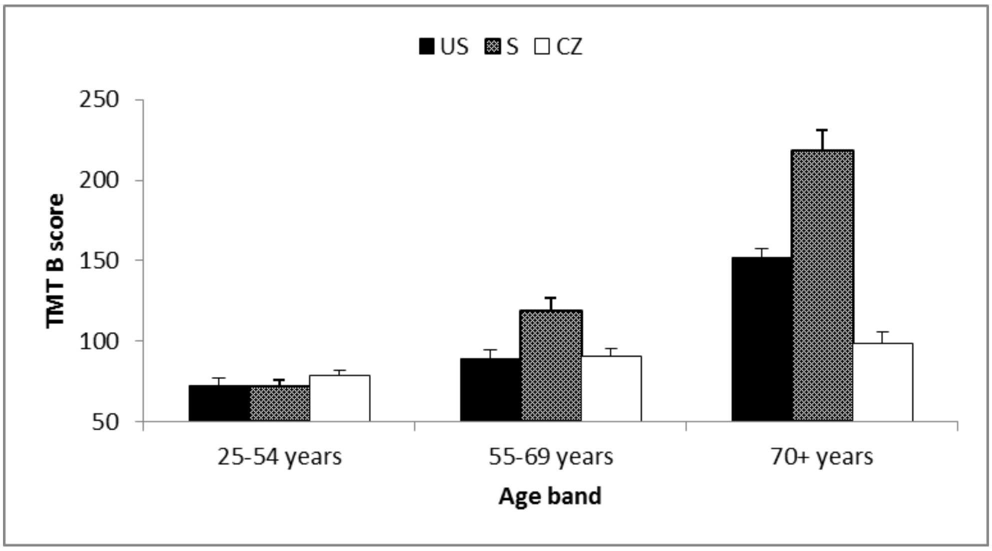

Figure 3. Reanalysis: Mean marginal estimates (education = 13.77 years) of TMT-B scores following different Sample x Age categories
Note. Groups: US, English language (EN) normative sample; S, Spanish normative sample; CZ, Czech normative sample. The bar graph suggests that there are pronounced differences in TMT-B performance between younger and older age groups and that these differences are significant even between EN and CZ samples. The Spanish sample should be treated with caution due to small sample size (N = 15 subjects) in 70+ age group.

Discussion

In this study, our goal was to determine the validity of English (North American; Schretlen et al, 2010), Czech (Bezdicek et al., 2012), and Spanish (Ojeda, Peña, unpublished data) normative data for the TMT, while controlling for as many potential biases as possible, e.g., differences in test construction (we used the same test format) and administration (all data according to Strauss, Sherman, & Spreen, 2006, p. 656).

As expected, we found that the English normative data from North America were not identical to those proposed recently in other sociocultural and language contexts (Czech and Spanish), namely in respect to TMT, Part B. These differences related especially to TMT, Part B, and more specifically to older Czech and English samples (70–74 years and older). We argue that the English TMT norms collected by Schretlen et al. (2010) should not be used in the Czech Republic or Spain (and perhaps in other European countries as well) unless these differences are taken into account. While the language differences observed in this study were largest between English and non-English languages, they were also found between two European countries (compare reanalysis on larger age groups). These findings further support the need for adjustment of available TMT norms for use in different languages and sociocultural contexts (Mitrushina, Boone, Razani, & D'Elia, 2005; Ojeda, Aretouli, Peña, & Schretlen, 2014). Ideally, pooling data from multiple countries and developing regression-based norms might prove to be the most effective means of accounting for these language models.

When the SDs in the first analysis were as large as 107 and 124 seconds, we tried to reanalyze our data with another sample that combined larger age groups. The reanalysis strengthened our case for language and sociocultural non-equivalence of the normative data sets. As regards statistical significance in TMT-A, it is of note that in the last age group, the S sample was relatively weak (N = 15). It could be argued that at least part of this result might be attributed to a small sample size of Spaniards (*N* = 15), however the difference observed between CZ and EN participants remains robust (*N*s = 53 and 88, respectively).

Arguably, the observed language-related differences in both TMT Part A and TMT Part B cannot be explained only by methodological biases. In fact, with the exception of one letter in the Czech version, all included data were based upon the same numeric and alphabetical system, thus eliminating sources of the construct bias other than the "language" bias itself as defined by Van de Vijver and Tanzer (2004, p. 124). In this respect, the stimulus familiarity can be expected to be equivalent across all languages, and instrument bias is negligible insofar as the alphabetical and numeric system belongs to common knowledge. The sample bias may also be considered to be eliminated, as both the sample sizes and education levels were comparable across all age groups. However, educational effects may go far beyond number of years of school attended and can be possibly a driving force in sociocultural differences in test performance (Cosentino, Manly, & Mungas, 2007; Ojeda, Aretouli, Peña & Schretlen, 2014). Finally, administration bias is excluded by the standard administration procedure.

This study has several limitations. First, the Spanish sample was younger and showed greater age-related differences in completion times than the other samples. Whether the latter reflects a greater prevalence of undiagnosed health problems with increasing age or cohort differences in education or some other factor cannot be determined based on the available data. Second, the Spanish sample includes relatively few subjects in the older age groups, with single digit cell sizes for those in the oldest four groups, compared to mostly double digit cell sizes for the same age groups in the EN and CZ samples. This likely contributes to the increased variability of standard deviations across age groups shown by the Spanish sample, although we tried to decrease this variability by dividing the samples into larger age groups. Third, all three studies were based on convenience samples rather than census-matched samples. Fourth, error scores are missing in the Czech normative data. These scores would allow for a more finely-tuned analysis of language differences, since although the Spanish and English data reveal slower TMT-B performance, it could be that Bezdicek et al. participants were prone to commit more errors which, in turn, would explain why their completion times might be shorter (e.g., we were unable in the present study to disentangle the potential confound of changes in the use of the "ch" digraph in the Czech alphabet that may lead to increased completion times due to scanning errors). Fifth, only Schretlen's et al. (2008) study mentioned the participants' ethnic background; it is likely that this English (North American) sample had a more diverse ethnic representation than the Czech and Spanish samples, where all participants were Caucasian. The question of using of "race-specific" norms is not without controversy (Pedraza & Mungaz, 2008) and remains open for further investigation through comparable normative studies in other European languages. Sixth, we did not control for the participants' health status apart from functional assessment (MMSE, history taking, and neuropsychological battery). Since only participants in the Schretlen et al. (2008) study underwent physical and neurological examination, it is possible that the general health status of participants differed across studies. Seventh, all individuals were seen without formal effort testing. Eighth, although the English (North American) sample took longer, on average (95.3 seconds), than the Czech and Spanish samples (81.6 seconds) to complete TMT-B, this was likely due to the greater representation of elderly participants in the English sample. Table 2 (appendix) shows that English TMT-B mean completion times were comparable to or faster than those shown by the Czech and Spanish participants in the seven youngest age cohorts and were intermediate between the Czech and Spanish participants in the oldest three cohorts. Finally, the differences between samples in this study as representative of nationality or language may overlook the enormous proportion of variance contributed by sampling artifact as seen in the literature, e.g., Drane, Yuspeh, Huthwaite, & Klingler (2002), the mean attenuated only for age is 153 seconds. Using the Mitrushina et al. (2005) meta-analytic tables (derived from studies performed “primarily” in the U.S. and Canada), the predicted (mean) score on TMT-B is about 116 seconds. From Tombaugh (2004) attenuated for age and education, it is 86 seconds. From Ashendorf et al. (2008) attenuated for age and education it is 81 seconds. So, the authors of the present study are confident that only a small nevertheless significant part of the discrepancy among our samples is language in origin.

Despite these limitations, our findings suggest that there is considerable variability in TMT norms across English (North American) and European samples, even after controlling for potential sources of bias. Normative data studies show first a lack of sociocultural equivalent, and second, a lack of language equivalence in TMT performance. One possible strategy to account for differences would be to pool normative data from different countries and then develop regression-based norms that include terms for the languages represented along with terms for age and other demographic determinants of test performance.

About the Authors

Ondrej Bezdicek, Ph.D. is with the Department of Neurology and Centre of Clinical Neuroscience, First Faculty of Medicine and General University Hospital in Prague, Charles University in Prague, Czech Republic. Correspondence concerning this article should be addressed to Ondrej Bezdicek, Ph.D., Department of Neurology and Centre of Clinical Neuroscience, First Faculty of Medicine and General University Hospital in Prague, Charles University in Prague, Kateřinská 30, 128 21 Praha 2, Czech Republic; +420 731 478 794 (office) ; +4202 2492 2678 (fax);
Email: ondrej.bezdicek@gmail.com (e-mail).

Ladislav Moták, Ph.D. is with the Université de Toulouse-Le Mirail, Cognition, Langues, Langage, Ergonomie – Laboratoire Travail et Cognition, UMR 5263 CNRS; Toulouse, France.

David J. Schretlen, Ph.D. is with the Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine; Baltimore, Maryland, USA.

Marek Preiss, Ph.D. is with the National Institute of Mental Health; Klecany, Czech Republic.

Bradley N. Axelrod, Ph.D. is with the John D. Dingell Department of Veterans Affairs Medical Center; Detroit, Michigan, USA.

Tomáš Nikolai, Ph.D. is with the Department of Neurology and Centre of Clinical Neuroscience, First Faculty of Medicine and General University Hospital in Prague, Charles University in Prague, Czech Republic.

Javier Peña, Ph.D. is with the Department of Methods and Experimental Psychology, Faculty of Psychology and Education, University of Deusto; Bilbao, Spain.

Natalia Ojeda, Ph.D. is with the Department of Methods and Experimental Psychology, Faculty of Psychology and Education, University of Deusto; Bilbao, Spain.

Evžen Růžička, D.Sc. is with the Department of Neurology and Centre of Clinical Neuroscience, First Faculty of Medicine and General University Hospital in Prague, Charles University in Prague, Czech Republic.

Disclosure

The work was supported by grants from the Czech Ministry of Education: PRVOUK-P26/LF1/4.

Acknowledgement

The authors are grateful to Claire Schretlen, B.S., for her help with editing the text for English language and grammar and gratefully acknowledge the contribution of anonymous reviewers.

References

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., ... Phelps C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia, 7*(3), 270–9.

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

Arsenault-Lapierre, G., Whitehead, V., Belleville, S., Massoud, F., Bergman, H., & Chertkow, H. (2011). Mild cognitive impairment subcategories depend on the source of norms. *Journal of Clinical and Experimental Neuropsychology, 33*(5), 596–603.

Ashendorf, L., Jefferson, A. L., O'Connor, M. K., Chaisson, C., Green, R. C., & Stern, R. A. (2008). Trail Making Test errors in normal aging, mild cognitive impairment, and dementia. *Archives of Clinical Neuropsychology, 23*, 129–37. doi: 10.1016/j.acn.2007.11.005. Epub 2008 Feb 21.

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.

Bezdicek, O., Motak, L., Axelrod, B. N., Preiss, M., Nikolai, T., Vyhnalek, M., ... Ruzicka, E. (2012) Czech version of the Trail Making Test: Normative data and clinical utility. *Archives of Clinical Neuropsychology, 27*, 906–914.

Cosentino, S., Manly, J, & Mungas, D. (2007). Do reading tests measure the same construct in multiethnic and multilingual older persons? *Journal of the International Neuropsychological Society*, *13*(2), 228–36.

Dickinson, M. D., Hiscock, M. (2011). The Flynn effect in neuropsychological assessment. *Applied Neuropsychology*, *18*(2), 136–42.

Dietz, A. (2011). Language. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of Clinical Neuropsychology* (pp. 1425–1426). New York: Springer.

Drane, D. L., Yuspeh, R. L., Huthwaite, J. S., & Klingler, L. K. (2002). Demographic characteristics and normative observations for derived-Trail Making Test indices. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology, 15*, 39–43.

Fernández, 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*, 239–46.

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

Hester, R. L., Kinsella, G. J., Ong, B., & McGregor, J. (2005). Demographic influences on baseline and derived scores from the Trail Making Test in healthy older Australian adults. *The Clinical Neuropsychologist, 19*, 45–54. doi:10.1080/13854040490524137

Lageman, S. K. (2011). Cultural Diversity in Neuropsychology. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of Clinical Neuropsychology* (pp. 752–754). New York: Springer.

Lamberty, G. J., Putnam, S. H., Chatel, D. M., Bielauskas, L. A., & Adams, K. M. (1994). Derived Trail Making Test indices: A preliminary report. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology, 7*, 230–234.

Litvan, I., Goldman, J. G., Tröster, A. I., Schmand, B. A., Weintraub, D. , Petersen, R. C., ... Emre, M. (2012). Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society Task Force guidelines. *Movement Disorders, 27*, 349–356.

Loewenstein, D. A., Argüelles, T., Argüelles, S., Linn-Fuentes, P. (1994). Potential cultural bias in the neuropsychological assessment of the older adult. *Journal of Clinical and Experimental Neuropsychology, 16*, 623–629.

Manly, J. J. (2008). Critical issues in cultural neuropsychology: Profit from diversity. *Neuropsychology Review, 8*, 179–183.

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of Normative Data for Neuropsychological Assessment*. (2nd ed.). New York, NY: Oxford University Press.

Ojeda, N., Aretouli, E., Peña, J., & Schretlen, D. J. (2014). Age differences in cognitive performance: A study of cultural differences in Historical Context. *Journal of Neuropsychology*. Advance online publication. doi: 10.1111/jnp.12059. \

Partington, J. E., & Leiter, R. G. (1949). Partington's Pathways Test. *Psychological Service Center Journal, 1*, 11–20.

Pedraza, O., Mungas, D. (2008). Measurement in cross-cultural neuropsychology. *Neuropsychology Review, 18*, 184–93.

Preiss, M., & Preiss, J. (2006). Test cesty. [Trail Making Test. Manual in Czech]. Bratislava: Psychodiagnostika.

Reitan, R. M. (1958). Validity of the Trail Making Test as an indication of organic brain damage. *Perceptual and Motor Skills, 8*, 271–276.

Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery* (2nd ed.). Tucson, AZ: Neuropsychology Press.

Schretlen, D. J., Munro, C. A., Anthony, J. C., & Pearlson, G. D. (2003). Examining the range of normal intraindividual variability in neuropsychological test performance. *Journal of the International Neuropsychological Society, 9,* 864–870. doi:10.1017/S1355617703960061

Schretlen, D. J., Testa, S. M., & Pearlson, G. D. (2010). *Calibrated Neuropsychological Normative System professional manual*. Lutz, Florida: Psychological Assessment Resources, Inc.

Schretlen, D. J., Testa, S. M., Winicki, J. M., Pearlson, G. D., & Gordon, B. (2008). Frequency and bases of abnormal performance by healthy adults on neuropsychological testing. *Journal of the International Neuropsychological Society, 14*(3), 436–445.

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

Strauss, E., Sherman, E., & Spreen, O. (2006). *A compendium of neuropsychological tests.* (3rd ed.). New York, NY: Oxford University Press.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson Education.

Testa, S. M., Winicki, J. M., Pearlson, G. D., Gordon, B., & Schretlen, D. J. (2009). Accounting for estimated IQ in neuropsychological test performance with regression-based techniques. *Journal of the International Neuropsychological Society*, *15*(6), 1012–1022.

Tombaugh, T. N. (2004). Trail Making Test A and B: Normative data stratified by age and education. *Archives of Clinical Neuropsychology, 19,* 203–14.

Urbánek, T. (2010). Nejpoužívanější psychodiagnostické metody v České republice. [The most used psychological assessment tools in the Czech Republic]. *Testfórum 1,* 6–9.

U.S. 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. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue européenne de psychologie appliquée, 54,* 119–135.

Appendix

Table 2

TMT performance in all three samples according to age bands and basic and derived TMT-indices

TMT-A

	EN			S			CZ		
Years	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
15–24	16	25.63	9.46	135	29.30	9.40	35	27.74	10.45
25–34	45	26.16	10.55	49	27.24	8.48	95	28.02	9.93
35–44	46	27.30	10.81	50	33.22	12.58	56	27.21	7.13
45–54	49	28.18	9.86	57	41.02	17.64	71	32.38	9.70
55–59	28	27.75	8.91	24	54.92	21.77	43	35.84	13.83
60–64	21	34.14	9.67	12	66.67	46.03	36	35.22	12.67
65–69	32	36.91	14.76	6	67.83	24.62	33	39.03	13.22
70–74	29	41.17	15.32	3	100.00	23.90	31	41.58	11.67
75–79	27	51.00	20.02	6	88.00	82.52	13	41.08	17.58
80+	32	58.66	20.37	6	121.83	46.13	9	54.11	26.36
Total	325	34.91	17.05	348	38.43	25.85	422	32.85	12.81

TMT-B

	EN			S			CZ		
Years	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
15–24	16	55.44	22.79	135	56.95	17.10	35	67.09	37.42
25–34	45	59.71	30.41	49	60.84	18.61	95	72.31	34.00
35–44	46	70.52	34.83	50	70.04	25.14	56	70.57	26.55
45–54	49	77.73	46.64	57	90.51	51.91	71	83.41	53.46
55–59	28	74.71	40.20	24	111.08	36.68	43	87.09	40.54
60–64	21	92.90	52.51	12	150.42	100.57	36	90.06	67.24
65–69	32	91.91	47.20	6	167.67	26.63	33	97.36	43.47
70–74	29	144.72	108.62	3	262.33	107.02	31	83.06	22.20
75–79	27	146.30	74.39	6	191.17	124.76	13	104.77	62.42
80+	32	163.34	99.17	6	277.67	91.86	9	133.89	48.62
Total	325	95.34	69.87	348	81.63	57.91	422	81.59	44.39

TMT B-A

	EN			S			CZ		
Years	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
15–24	16	29.81	22.02	135	27.65	14.83	35	39.34	30.60
25–34	45	33.56	24.55	49	33.59	14.31	95	44.28	29.88
35–44	46	43.22	29.36	50	36.82	23.39	56	43.36	25.21
45–54	49	49.55	43.70	57	49.49	40.98	71	51.03	47.67
55–59	28	46.96	34.77	24	56.17	21.80	43	51.26	32.87
60–64	21	58.76	47.59	12	83.75	60.09	36	54.83	59.87
65–69	32	55.00	39.17	6	99.83	15.33	33	58.33	39.92
70–74	29	103.55	102.08	3	162.33	83.76	31	41.48	20.92
75–79	27	95.30	70.20	6	103.17	54.02	13	63.69	48.10
80+	32	104.69	89.40	6	155.83	66.35	9	79.78	47.45
Total	325	60.42	60.34	348	43.20	37.48	422	48.74	38.23

TMT B/A

	EN			S			CZ		
Years	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
15–24	16	2.30	0.77	135	2.04	0.64	35	2.41	0.81
25–34	45	2.36	0.98	49	2.32	0.64	95	2.66	1.00
35–44	46	2.62	0.88	50	2.29	0.97	56	2.67	1.01
45–54	49	2.87	1.33	57	2.24	0.82	71	2.53	1.02
55–59	28	2.65	0.86	24	2.10	0.48	43	2.49	0.93
60–64	21	2.70	1.07	12	2.28	0.43	36	2.51	0.99
65–69	32	2.56	1.14	6	2.63	0.60	33	2.61	1.08
70–74	29	3.58	2.22	3	2.55	0.53	31	2.09	0.60
75–79	27	3.06	1.55	6	2.52	0.93	13	2.49	0.70
80+	32	2.82	1.16	6	2.36	0.65	9	2.70	0.86
Total	325	2.75	1.28	348	2.19	0.72	422	2.54	0.96

Note. Groups: EN, English normative sample; S, Spanish normative sample; CZ, Czech normative sample; TMT, Trail Making Test; M, mean; SD, standard deviation; TMT-A and TMT-B completion times in seconds, derived indices: TMT (B - A) difference score and the TMT (B/A) ratio score.