Journal of the International Neuropsychological Society (2023), **29**, 783–788
doi:10.1017/S1355617722000911

Research Article

Parallel and identical test–retest reliability of the Tower of London test – Freiburg version

Valentin Schyle, Lena V. Schumacher, Benjamin Rahm and Josef M. Unterrainer
Institute of Medical Psychology and Medical Sociology, Faculty of Medicine, University of Freiburg, Germany

Abstract

Objectives: The Tower of London – Freiburg version (TOL-F) was developed in three parallel-test versions (A, B, and C) that only differ in their physical appearance by interchanged ball colors, but not in their cognitive demands. We addressed the question whether the test–retest reliability of an identical problem set differs from the parallel test–retest reliability of a structurally identical problem set with a marginally different physical appearance. **Methods:** Reliabilities were assessed in two samples of young adults over a 1-week interval: In the parallel test–retest sample ($n = 93$; 49 female), half of the participants accomplished version A at the first session and version B at the second session, while the other half started with version B in the first session and continued with A in the second session. In the identical test–retest sample ($n = 86$; 48 female), half of the participants performed on version A in both the first and the second session, while the other half went through the same procedure with version B. **Results:** For overall planning accuracy, intraclass correlation coefficients for absolute agreement were $r = .501$ for the parallel test–retest and $r = .605$ for the identical test–retest sample, with Pearson correlations of $r = .559$ and $r = .708$ respectively. Greatest lower bound estimates of reliability were adequate to high in the two samples (ranging between .765 and .854) confirming previous studies. **Conclusions:** Although the TOL-F revealed only moderate intraclass correlations for absolute agreement, it showed some of the highest psychometric indices compared to repeated assessments with other TOL tests.

Keywords: Tower of London Test – Freiburg version; parallel test–retest reliability; identical test–retest reliability; intraclass correlation coefficients

(Received 3 August 2022; final revision 6 October 2022; accepted 8 November 2022; First Published online 12 December 2022)

Introduction

When Tim Shallice first introduced the Tower of London (TOL) planning task to measure frontal brain functions (Shallice, 1982), this was the starting point for a series of further developments of variants and versions of the TOL and other so-called disk-transfer tasks (Berg & Byrd, 2002). One reason for these diverse developments was the rather unsatisfactory reliability of the original TOL version in adults (Cronbach's $\alpha = .25$, split-half reliability $r = .19$, Humes, Welsh, & Retzlaff, 1997; see also Michalec et al., 2017; test–retest reliability, $r = .60$; Lowe & Rabbitt, 1998), and in children (Cronbach's α could not be determined, test–retest reliability was $r = .23$; Syväoja et al., 2015). Today, there are several versions and variants of the TOL task that feature acceptable psychometric properties (e.g., Culbertson & Zillmer, 2005: test–retest $r = .75$ for total moves, $r = .59$ for total correct in patients with Parkinson's disease; Schnirman, Welsh, & Retzlaff, 1998: test–retest $r = .70$; Tucha, & Lange, 2004: test–retest $r = .85$; Unterrainer et al., 2019: internal consistency $glb = .76$).

In the context of cognitive tasks, reliability indexes the stability of measurements and features mainly two aspects: (i) the task's internal consistency and split-half reliability reflect the degree to which all items of the task measure the same underlying construct and (ii) the consistency between repeated measurements of identical (test–retest reliability) or highly similar versions (parallel test–retest reliability) of the task. In the present study, we will focus on the latter aspect by studying the test–retest and parallel test–retest reliability of the TOL. Previous studies have mainly reported the Pearson correlation coefficient. However, it is no longer considered an ideal measure of identical and parallel test–retest reliability, as it only captures the relative consistency but not the absolute agreement of test scores over repeated measurements. For absolute agreement, total score variance is taken into account, including not only the variance between two measurements but also within the sample (McCraw & Wong, 1996). There is a growing consensus towards the use of different forms of the intraclass correlation coefficient (ICC, see Shrout & Fleiss, 1979; McCraw & Wong, 1996), as these may inform about both the relative consistency [ICC(3,1)] and the absolute agreement [ICC(2,1)] between the repeated measurements.

Tyburski, Kerestey, Kerestey, Radoń, & Mueller (2021) have recently provided a comprehensive overview of identical and parallel test–retest reliability studies of TOL versions, which also lists four studies on adults that reported ICCs. It is noticeable that

Corresponding author: Josef M. Unterrainer, email: josef.unterrainer@mps.uni-freiburg.de
Cite this article: Schyle V., Schumacher L.V., Rahm B., & Unterrainer J.M. (2023) Parallel and identical test–retest reliability of the Tower of London test – Freiburg version. *Journal of the International Neuropsychological Society*, **29**: 783–788, https://doi.org/10.1017/S1355617722000911

Table 1. Descriptive and inferential statistics for sample characteristics of demographic information, and scores on tests of depressive symptoms, fluid and crystallized intelligence

					Statistical comparisons	
	Parallel test–retest sample		Identical test–retest sample		Statistical test	*p*-values
Sex	Male = 44	Female = 49	Male = 38	Female = 48	Chi-square 0.176	.675
Age	Mean 21.93	SD 1.95	Mean 22.01	SD 2.33	T-value −0.258	.797
BDI-II	Mean 4.33	SD 3.45	Mean 4.59	SD 3.72	T-value −0.485	.628
MWT-B	Mean 25.59	SD 3.92	Mean 24.98	SD 4.19	T-value 1.016	.311
APM	Mean 11.40	SD .90	Mean 11.20	SD 1.12	T-value 1.316	.190

with the exception of one study (Köstering, Nitschke, Schumacher, Weiller, & Kaller, 2015; ICC(2,1) = 0.69 for accuracy in terms of total optimal solutions), the ICCs for the performance parameters remained considerably below the desired requirements of at least 0.5, indicating poor reliability (Portney & Watkins, 2000). More specifically, for outcomes that consider the number of problems solved, Lemay, Bédard, Rouleau, & Tremblay (2004) report an ICC(2,1) of 0.33, Tunstall, O'Gorman, & Shum (2016) an ICC(2,1) of 0.45, and Tyburski et al. (2021) even a negative ICC(3,1) of −0.44.

This observation of low replicability of TOL measurements is neither new nor surprising when seen in the wider context of similar findings for other tasks measuring higher-order executive functions (Burgess, 1997; Rabbitt, 1997). Planning as a prototypical higher-order executive function reflects the mental generation and evaluation of potential solution alternatives in novel problem situations. This novelty aspect particularly hampers the test–retest reliability assessment of planning tasks, as novelty is not given in a second measurement with identical problem items, and practice effects are likely to occur (Rabbitt, 1997; Strauss, Sherman, & Spreen, 2006). One way to avoid using identical items for the second measurement is to use an alternative or parallel-test version. Accordingly, Calamia, Markon, & Tranel (2012) observed that the use of alternate forms helps to decrease the size of practice effects, although it does not necessarily increase reliability. In a meta-analysis of test–retest correlations of instruments typically used in neuropsychological assessment (Calamia, Markon, & Tranel, 2013), for a majority of tests the application of a parallel form at retesting was associated with a decrease in the test–retest correlation in comparison to retesting the identical form. Although the magnitude of the differences was generally $\Delta r = .1$ or less, according to the authors, psychometric properties like difficulty can differ significantly between test versions.

In this respect, it is an open question whether the test–retest reliability of an identical problem set differs from the parallel test–retest reliability of an alternative but highly similar problem set. One instrument that could be used to systematically address this question is the TOL-Freiburg version (Kaller, Unterrainer, Kaiser, Weisbrod, & Aschenbrenner, 2012a). This planning test has a sufficiently high internal consistency ($glb = .73$ and .76, Kaller et al. 2016, Unterrainer et al. 2019, respectively) and hence fulfills basic psychometric requirements. It was developed in three parallel-test versions (A, B, and C), whose problems are identical in structure, but whose physical appearance differs due to permutations of ball colors. Thus, these versions should require identical cognitive demands since structural problem parameters like search depth and goal hierarchy were kept completely identical (Kaller, Rahm, Köstering, & Unterrainer, 2011; Kaller, Unterrainer, & Stahl, 2012b). Köstering et al. (2015) already assessed test–retest reliability of the TOL-Freiburg using version A at the first and B at the second session over a 1-week interval. Pearson correlation ($r = .739$) and ICC for absolute agreement ($r = .690$) yielded adequate test–retest reliabilities. As in this study participants performed two different versions and the sample size war rather small (n = 27), here we addressed the question whether the test–retest reliability of an identical problem set (versions A-A and B-B) differs from the parallel test–retest reliability (versions A-B and B-A) on the basis of two larger samples.

Methods

Participants

The study comprised two separate, completely non-overlapping samples including only participants who had no previous experience with the TOL test.

For the parallel test–retest sample, we originally recruited 103 young participants with predominantly high school degrees or who are studying. Inclusion criteria were sufficient German language skills to ensure comprehension of task instructions, age between 18 and 26 years, and normal or corrected-to-normal vision. Exclusion criteria were current/past psychiatric or neurological disease, psychotropic medication, and color blindness. Depressive symptoms, crystallized, and fluid intelligence were assessed with the Beck Depression Inventory-II (BDI-II; Beck, Steer, & Brown, 1996), a German vocabulary test (Mehrfachwahl-Wortschatz-Intelligenztest or MWT-B; Lehrl, 2005), and with the Advanced Progressive Matrices (short version, German adaptation and norming, Bulheller, & Häcker, 1998), respectively. Due to increased depression scores (BDI score above 14), ten subjects had to be excluded. The final parallel test–retest sample (N = 93; 49 females) had a mean age of 21.9 years (SD = 1.95; range 18.33–25.92). Participants were compensated with 20€ for both sessions. In the parallel test–retest sample, half of the participants accomplished version A at the first session and version B at the second session, while the other half started with version B in the first session and continued with A in the second session.

For the identical test–retest sample, 93 young participants with predominantly high school degrees or who are studying were recruited applying identical inclusion/exclusion criteria, screening for depressive symptoms, crystallized and fluid intelligence tests, and compensation as for the parallel test–retest sample. In consequence, seven participants had to be excluded resulting in the final identical test–retest sample of 86 participants (48 female) with a mean age of 22.01 (SD = 2.32; range 18.08–26.42). In the identical test–retest sample, half of the participants performed version A in both the first and the second session, while the other half went through the same procedure with version B. Table 1 provides a comparative overview of both samples.

The study was approved by local ethics authorities (EK-Freiburg nr. 479/19). Data acquisition complied with local institutional research standards for human research and was completed in accordance with the Helsinki Declaration.

https://doi.org/10.1017/S1355617722000911 Published online by Cambridge University Press

Tower of London – Freiburg Version (TOL-F)

All participants were tested individually in a quiet room with the TOL-F (Kaller et al., 2012a). The TOL-F is as a computerized pseudo-realistic representation of the TOL's originally wooden configuration and is implemented in the Vienna Test System (https://marketplace.schuhfried.com/de/tol). Individual problem items consist of a start and a goal state that are presented in the lower and upper halves of the computer screen, respectively. In order to transfer the start into the goal state, the TOL-F can be worked on by touch screen. Thus, a ball is picked up simply by clicking the ball via finger touch. The selected ball is then encircled by a transparent whitish corona and can be moved by selecting the respective rod by finger touch.

Subjects are instructed to transform the start state into the goal state in the minimum number of moves which are shown to the left of the start state. Written instructions inform that only one ball may be moved at a time, that balls cannot be placed beside the rods, that only the top-most ball can be moved in case several balls are stacked on a rod, and that the rods differ in their capacities of accommodating one, two, or three balls at maximum. The computer program does not allow breaking these rules, but records any attempts to do so. Instructions further emphasize that problems have to be solved in the minimum number of moves and that participants should always plan ahead the problem solution before starting with movement execution.

For assessment of individual planning ability with the TOL-F, overall planning accuracy, defined as the sum of problems that were correctly solved in the minimum number of moves, is regarded as the primary outcome variable of interest. The TOL-F provides three different levels of minimum moves (four, five, and six move problems, eight of each, presented in increasing minimum number of moves) resulting in an overall planning accuracy of 24 problems at maximum (corresponding to 100 percent). A one-minute time limit per trial was implemented, as in the original study of Shallice (1982).

The TOL-F features three parallel-test versions, A, B, and C, consisting of the same set of problems, which are color-permuted; that is, ball colors are interchanged (cf. Berg & Byrd, 2002). Thus, across parallel-test versions, problems are structurally identical, while their physical appearance is different. As already described in the Participants section, in the parallel test–retest sample, half of the participants accomplished version A at the first session and version B at the second session, while the other half started with version B in the first session and continued with A in the second session. In the identical test–retest sample, half of the participants performed version A in both the first and the second session, while the other half went through the same procedure with version B.

Analyses

First, to compare planning accuracy between the two samples and to assess changes over the two time points, a repeated-measurements ANOVA (RM-ANOVA) was calculated with the within-subject factor session (1 versus 2) and the between-subjects factor group (parallel test–retest versus identical test–retest). For assessing parallel and identical test–retest reliabilities, ICCs using two-way random effects models of absolute agreement ICC(2,1) and relative consistency ICC(3,1) corresponding to Shrout and Fleiss (1979) were computed. For comparability with previous studies, we also report Pearson product–moment correlations as indices of identical/parallel test–retest reliability as well as glb (estimations of greatest lower bound) as index of internal consistency.

Results

Overall planning performance

RM-ANOVA with the within-subject factor session (1 versus 2) and the between-subjects factor group (parallel versus identical test–retest) on planning accuracy revealed significant main effects for session ($F(1, 177) = 70.010$, $p = <.001$; $\eta^2_{partial} = .283$) and group ($F(1, 177) = 6.076$, $p = .015$; $\eta^2_{partial} = .033$), but no interaction effect ($F(1, 177) = 1.175$, $p = .280$; $\eta^2_{partial} = .007$). As obvious from Figure 1 and descriptive data of Table 2, participants increased planning performance on average about 6.5% (i.e., about 1.5 out of 24 problem items) from the first to the second session. In addition, the parallel test–retest group performed about 5% better than the identical test–retest group across both sessions.

To additionally check whether the order with which version testing has started is associated with a different learning process, the analysis above is carried out separately for both samples, but supplemented with the between-subject factor start (A versus B).

For the parallel test–retest sample, there was again a significant main effect for session ($F(1, 91) = 24.056$, $p = <.001$; $\eta^2_{partial} = .209$), but not for the factor start ($F(1, 91) = .877$, $p = .351$; $\eta^2_{partial} = .010$) or the interaction effect ($F(1, 91) = 1.714$, $p = .194$; $\eta^2_{partial} = .018$).

In the identical test–retest sample, there was also a significant main effect for session ($F(1, 84) = 50.595$, $p = <.001$; $\eta^2_{partial} = .376$), but not for start ($F(1, 84) = 0.013$, $p = .911$; $\eta^2_{partial} = .000$), or the interaction effect ($F(1, 84) = 1.298$, $p = .258$; $\eta^2_{partial} = .015$). In both samples performance increased from the first to the second session, but there was no difference between starting with version A versus B or an interaction with learning across repeated assessments.

Internal consistency (glbs)

The greatest lower bound estimations for the parallel test–retest sample were 0.765 and 0.854 for session 1 and session 2, respectively. In the identical test–retest sample, glbs were 0.806 and 0.817 for session 1 and 2, respectively.

Parallel test–retest and identical test–retest reliability

In the parallel test–retest sample, overall planning accuracy for repeated assessments with different versions (A-B and B-A) showed a Pearson correlation of $r = .559$ (95% confidence interval [.400, .684], $p = .001$), a relative consistency ICC(3,1) of $r = .556$ (95% CI [.398, .682], $p = .001$), and an absolute agreement ICC(2,1) of $r = .501$ (95% CI [.268, .664], $p = .001$).

In the identical test–retest sample, overall planning accuracy for repeated tests with identical versions (A-A and B-B) revealed a Pearson correlation of $r = .708$ (95% CI [.584, .800], $p = .001$), a relative consistency ICC(3,1) of $r = .708$ (95% CI [.585, .800], $p = .001$), and an absolute agreement ICC(2,1) of $r = .605$ (95% CI [.204, .791], $p = .001$).

To check whether numerically higher reliability in the identical test–retest sample may be related to differences in variance, we also compared the variance of the overall performance between groups with Levene's test. As a result, sessions did not differ significantly, in line with assumed equality of variance (Session 1: $F(1, 177) = 0.378$; $p = .539$; Session 2: $F(1, 177) = 0.000$; $p = .991$). This was also true for the difference between sessions: According to Levene's test, there was no significant difference between group variances with regard to this difference ($F(1, 177) = 2.119$; $p = .147$).

https://doi.org/10.1017/S1355617722000911 Published online by Cambridge University Press

Table 2. Descriptive statistics of the TOL-F for overall planning accuracy

	Parallel test–retest sample			Identical test–retest sample		
	Mean	Range (Min–Max)	SD	Mean	Range (Min–Max)	SD
Session 1	75.18	37.50 to 100.00	11.51	70.16	33.33 to 95.83	12.73
Session 2	81.00	37.50 to 100.00	12.78	77.71	41.67 to 100.00	12.89
Difference	5.82	−29.17 to 29.17	11.46	7.56	−20.83 to 29.17	9.79

Note. Min = minimum; Max = maximum; SD = standard deviation; Difference score in accuracy is computed by subtracting Session 1 from Session 2.

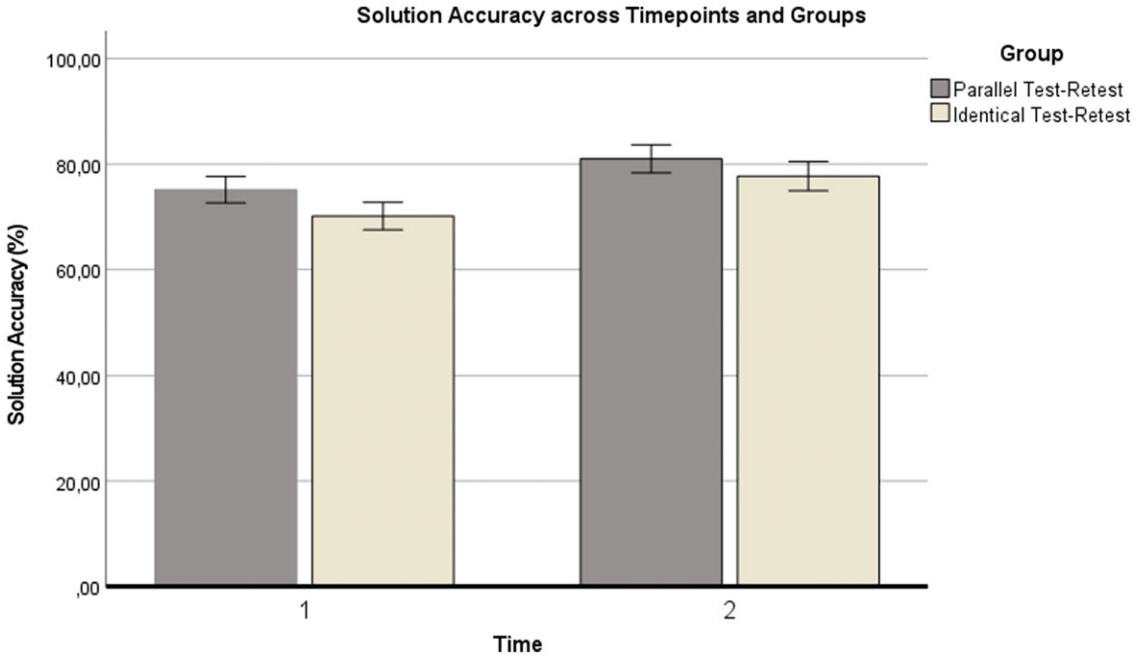

Figure 1. Overall planning accuracy in percent across sessions 1 and 2 for both samples (gray bars, parallel test–retest sample; beige bars, identical test–retest sample), with error bars denoting the standard error of mean.

Discussion

This study examined parallel and identical test–retest reliability of the TOL-F and revealed the following results: On the one hand, reliability was numerically higher for repeated assessment with the identical version compared to the parallel-test version. On the other hand, we found higher ICC absolute agreement measures than in most previous TOL studies. Except for the study by Köstering et al. (2015), no ICC(2,1) values for absolute agreement above .45 have been published so far for any TOL version (Tyburski et al., 2021). With ICCs(2,1) of .501 and .605 for parallel test–retest and identical test–retest reliability, we obviously exceed these values. For both results, however, it must be noted that the confidence intervals in the current study are rather large. Thus, the range of the true reliability value between both the parallel and the identical test–retest version and in comparison with previous studies does not indicate a significant difference.

The main question of this study was the comparison of retesting an identical versus an alternative version. In line with the results of Calamia et al. (2013), the identical versions achieved numerically higher reliability than alternative versions. Calamia et al. call for an ideally psychometrically identical alternative version, although this is not the case for many measurements (Calamia et al., 2012). TOL-F versions A and B consist of the same set of problems, only the ball colors are interchanged (cf. Berg & Byrd, 2002). Thus, we concluded that the color permutation should correspond to the idea of an ideal parallel version and at least reduce item-specific learning. General learning of the task remains, but that should always be the case. Numerically, it seems that the exchange of colors can lead to different reliabilities. Nevertheless, this conclusion is restricted by the overall performance difference between the two groups. It was confirmed again that the TOL-F problem set, that is balanced according to known structural problem parameters (Kaller et al., 2011), can exhibit satisfactory psychometric properties and even exceed internal consistencies established earlier (Kaller et al. 2012b, 2016, Unterrainer et al., 2019). However, the present ICC values can only be rated as "moderate" (ranged between .5 and .75) according to the criteria of Portney and Watkins (2000). This probably reflects the double-faced nature of executive functions and reliability. In their meta-analysis of some of the most widely used neuropsychological tests, Calamia et al. (2013) demonstrated that EF tests had poorer test–retest reliabilities compared to other measures of cognitive performance ($r < .70$). One explanation was the assumption that complex EF tasks involve multiple cognitive processes which makes them more susceptible to performance variability in repeated testing (Delis, Kramer, Kaplan, & Holdnak, 2004). In other words, the intended measurements of higher-order cognitive functions such as planning can also be strongly influenced by basic ongoing processes such as attention or working memory. Another explanation for lower reliability could be a learning effect that affects the second measurement: According to Strauss et al. (2006), practice effects

https://doi.org/10.1017/S1355617722000911 Published online by Cambridge University Press

on EF tests can lead to a restriction of range in test scores which in turn result in lower test–retest correlations. However, this assumption is only partly consistent with the present data and the analyses of Calamia et al. (2012, 2013). Probably it is not the size of the practice effect but individual changes in the rankings between the two measurement points that explain the different reliabilities (individual change of position in the second measurement, Duff, 2012). In very homogeneous samples as in our study the range in test scores is more restricted than in representative samples of the population (Strauss et al. 2006). Participants of the same age with similar cognitive abilities suggest less variance in performance than a more heterogeneous group with large age and educational differences. Lower variances render the same ranking in the second measurement less likely and thus may also lead to lower reliability scores.

But how can the noticeably higher ICCs of about $\Delta r = .2$ in the study by Köstering et al. (2015) be explained? After all, this study used the same TOL version as in the current parallel test–retest sample (A-B versus B-A), the test interval was identical, and the participants were students of the same age with similar intelligence scores and were recruited using the same exclusion and inclusion criteria. Apart from random sample variance, the extreme value adjustment in Köstering et al. may be an explanation. Since they studied a small sample ($n = 29$), they had to omit two cases deviating more than 2.5 standard deviations from the mean z-standardized between-session difference score to obtain reasonably normally distributed data. The two outliers were at the negative end of the distribution. This means that their performance in the second measurement was in the opposite direction of the whole group, which showed better performance in the second measurement. Duff (2012) has described how impressively test–retest reliabilities decrease when second measurements go in the contrary direction. The sample in the present study, which was more than three times larger, produced an acceptable normal distribution of the data per se, so that all values at both ends of the scale were included.

Limitations

A clear constraint is the rather homogeneous sample. A broader sample in terms of age and education would presumably allow the reliabilities to be increased even further and would offer better generalizability to the population. In addition, the recording of patient groups would be desirable. Although in both studies a total of more than 180 subjects were tested, the sample size is still below Watson's (2004) recommendation of at least 300 participants. In order to better quantify learning effects, several retests with different time intervals should be conducted. The overall performance difference between the two groups was an undesired outcome and might be related to the time period of data collection. While the parallel test–retest assessment was finalized before the Corona pandemic, the identical test–retest reliability measurements took place during the pandemic. Testing conditions therefore were slightly different due to the need to wear a face mask and to keep a greater interpersonal distance. In addition, one may speculate that due to reduced social contact and suspended face-to-face teaching, students may have been in a generally poorer mental and emotional condition during this time.

Conclusion

Even though the reliabilities obtained were only moderate, as can commonly be observed with EF, the present study showed some of the highest psychometrics for the TOL test. The small difference in reliability values between identical and parallel versions speak in favor of using the same version, as this allows us to expect more stable results over two measurement points.

Acknowledgements. This study was not funded by any third-party public funders, foundations, or companies, but exclusively with in-house resources.

Conflicts of interest. JMU declares to receive a small proportion of the license fees for the Freiburg version of the Tower of London (TOL-F) task from the SCHUHFRIED GmbH due to authorship of the published test materials (Kaller, Unterrainer, Kaiser, et al., 2012a).

References

Beck, A., Steer, R., & Brown, G. (1996). *Manual for the beck depression inventory–II*. The Psychological Corporation, Harcourt Brace.

Berg, W.K., & Byrd, D. L. (2002). The Tower of London spatial problem-solving task: Enhancing clinical and research implementation. *Journal of Clinical and Experimental Neuropsychology*, *24*, 586–604.

Bulheller, S., & Häcker, H. O. (1998). (Hrsg.): Advanced Progressive Matrices (APM). Deutsche Bearbeitung und Normierung nach J. C. Raven. Pearson Assessment.

Burgess, P. W. (1997). Theory and methodology in executive function research. In Methodology of frontal and executive function. Psychology Press. pp. 81–116. https://doi.org/10.4324/9780203344187-8

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *Clinical Neuropsychologist*, *26*, 543–570. https://doi.org/10.1080/13854046.2012.680913

Calamia, M., Markon, K., & Tranel, D. (2013). The robust reliability of neuropsychological measures: Meta-analyses of test-retest correlations. *Clinical Neuropsychologist*, *27*, 1077–1105. https://doi.org/10.1080/13854046.2013.809795

Culbertson, W., & Zillmer, E. (2005). The Tower of London-Drexel University (TOLDX^{TM}). Technical Manual, 2nd Edition. Multi-Health Systems.

Delis DC, Kramer JH, Kaplan E, Holdnack J. Reliability and validity of the Delis-Kaplan executive function system: An update. (2004) *Journal of the International Neuropsychological Society*, *10*, 301–303. https://doi.org/10.1017/S1355617704102191.

Duff, K. (2012). Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology*, *27*, 248–261. http://dx.doi.org/10.1093/arclin/acr120

Humes, G., Welsh, M., & Retzlaff, P. (1997). Towers of Hanoi and London: Reliability and validity of two executive function tasks. *Assessment*, *4*, 249–257.

Kaller, C.P., Debelak, R., Köstering, L., Egle, J., Rahm, B., Wild, P.S., & Unterrainer, J.M. (2016). Assessing Planning Ability Across the Adult Life Span: Population-Representative and Age-Adjusted Reliability Estimates for the Tower of London (TOL-F). *Archives of Clinical Neuropsychology*, *31*, 148–164.

Kaller, C. P., Rahm, B., Köstering, L., & Unterrainer, J. M. (2011). Reviewing the impact of problem structure on planning: A software tool for analyzing tower tasks. *Behavioural Brain Research*, *216*, 1–8.

Kaller, C. P., Unterrainer, J. M., Kaiser, S., Weisbrod, M., & Aschenbrenner, S. (2012a). *Tower of London—Freiburg version*. Schuhfried.

Kaller, C. P., Unterrainer, J. M., & Stahl, C. (2012b). Assessing planning ability with the Tower of London task: Psychometric properties of a structurally balanced problem set. *Psychological Assessment*, *24*, 46–53.

Köstering, L., Nitschke, K., Schumacher, F.K., Weiller, C., & Kaller, C.P. (2015). Test-retest reliability of the Tower of London Planning Task (TOL-F). *Psychological Assessment*, *27*, 925–931.

Lehrl, S. (2005). Mehrfachwahl-Wortschatz-Intelligenztest MWT-B. Spitta Verlag.

Lemay, S., Bédard, M. A., Rouleau, I., & Tremblay, P. L. G. (2004). Practice effect and test-retest reliability of attentional and executive tests in middle-aged to elderly subjects. *Clinical Neuropsychologist*, *18*, 284–302. https://doi.org/10.1080/13854040490501718

Lowe, C., & Rabbitt, P. (1998). Test\re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: theoretical and practical issues. Neuropsychologia, 36, 915–923.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Michalec, J., Bezdicek, O., Nikolai, T., Harsa, P., Jech, R., Silhan, P., Hyza, M., Ruzicka, E., & Shallice T. (2017). A comparative study of tower of London scoring systems and normative data. Archives of Clinical Neuropsychology, 32, 328–338. https://doi.org/10.1093/arclin/acw111

Portney LG, & Watkins MP. (2000). *Foundations of clinical research: Applications to practice*. Prentice Hall.

Rabbitt, P. (1997). Introduction: Methodologies and models in the study of executive functions. In Methodology of frontal and executive function. Psychology Press. pp. 1–38.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Schnirman, G.M., Welsh, M.C., & Retzlaff, P.D. (1998). Development of the Tower of London-Revised. *Assessment*, *5*(4), 355–360. doi: 10.1177/107319119800500404

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *298*, 199–209.

Strauss, E., Sherman, E., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*, 3rd Edition. Oxford University Press.

Syväoja, H.J., Tammelin, T.H., Ahonen, T., Räsänen, P., Tolvanen, A., & Kankaanpää, A., (2015). Internal consistency and stability of the CANTAB neuropsychological test battery in children. *Psychological Assessment*, *27*, 698–709.

Tyburski, E., Kerestey, M., Kerestey, P., Radoń, S., & Mueller, S. T. (2021). Assessment of motor planning and inhibition performance in non-clinical sample—reliability and factor structure of the tower of London and go/no go computerized tasks. *Brain Sciences*, *11*, 1420. https://doi.org/10.3390/brainsci11111420

Tucha, O, & Lange, W. (2004). TL-D. Turm von London – Deutsche Version. Hogrefe

Tunstall, J. R., O'Gorman, J. G., &Shum, D. H. K. (2016). A four-disc version of the Tower of London for clinical use. *Journal of Neuropsychology*, *10*, 116–129. https://doi.org/10.1111/jnp.12060.

Unterrainer, J. M., Rahm, B., Kaller, C. P., Wild, P. S., Münzel, T., Blettner, M., & Beutel, M. E. (2019). Assessing planning ability across the adult life span in a large population-representative sample: reliability estimates and normative data for the Tower of London (TOL-F) task. Journal of the International Neuropsychological Society, 1–10. https://doi.org/10.1017/S1355617718001248

Unterrainer, J.M., Rauh, R., Rahm, B., Hardt, J., Kaller, C.P., Klein, C., & Biscaldi, M. (2016). Development of planning in children with high-functioning autism spectrum disorders and/or attention deficit/hyperactivity disorder. *Autism Research*, *7*, 739–751.

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, *38*, 319–350. https://doi.org/10.1016/j.jrp.2004.03.001

https://doi.org/10.1017/S1355617722000911 Published online by Cambridge University Press