Assessing Planning Ability Across the Adult Life Span: Population-Representative and Age-Adjusted Reliability Estimates for the Tower of London (TOL-F)

Christoph P. Kaller1,2,3,*, Rudolf Debelak4, Lena Köstinger1,2,3,5, Johanna Egle4, Benjamin Rahm6, Philipp S. Wild7,8,9, Maria Blettner10, Manfred E. Beutel11, Josef M. Unterrainer6

1Department of Neurology, University Medical Center Freiburg, Freiburg, Germany
2Freiburg Brain Imaging Center, University of Freiburg, Freiburg, Germany
3BrainLinks-BrainTools Cluster of Excellence, University of Freiburg, Freiburg, Germany
4SCHUHFRIED GmbH, Mödling, Austria
5Biological and Personality Psychology, Department of Psychology, University of Freiburg, Freiburg, Germany
6Medical Psychology and Medical Sociology, University Medical Center Mainz, Mainz, Germany
7Preventive Cardiology and Preventive Medicine, Department of Medicine II, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany
8German Center for Cardiovascular Research (DZHK), partner site Hamburg/Kiel/Lübeck, Hamburg, Germany
9German Center for Cardiovascular Research (DZHK), partner site RhineMain, Mainz, Germany
10Institute of Medical Biostatistics, Epidemiology and Informatics, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany
11Department of Psychosomatic Medicine and Psychotherapy, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany

*Corresponding author at: Department of Neurology, University Medical Center Freiburg, Breisacher Str. 64, 79106 Freiburg, Germany.
Tel.: +49-761-270-54782; fax: +49-761-270-54819.
E-mail address: christoph.kaller@uniklinik-freiburg.de (C. Kaller).

Accepted 20 November 2015

Abstract

Planning ahead the consequences of future actions is a prototypical executive function. In clinical and experimental neuropsychology, discount-transfer tasks like the Tower of London (TOL) are commonly used for the assessment of planning ability. Previous psychometric evaluations have, however, yielded a poor reliability of measuring planning performance with the TOL. Based on theory-grounded task analyses and a systematic problem selection, the computerized TOL-Freiburg version (TOL-F) was developed to improve the task’s psychometric properties for diagnostic applications. Here, we report reliability estimates for the TOL-F from two large samples collected in Mainz, Germany (n = 3,770; 40–80 years) and in Vienna, Austria (n = 830; 16–84 years). Results show that planning accuracy on the TOL-F possesses an adequate internal consistency and split-half reliability (>0.7) that are stable across the adult life span while the TOL-F covers a broad range of graded difficulty even in healthy adults, making it suitable for both research and clinical application.

Keywords: Tower of London; Planning; Problem solving; Reliability; TOL-F

Introduction

More than 30 years ago, the Tower of London (TOL) planning paradigm was introduced as a prototypical task for assessing executive functions (EF) and the control of behavior in nonroutine situations (Shallice, 1982). Impairments of EF occur in various clinical populations and significantly interfere with patients’ everyday functioning (Chan, Shum, Touloupoulou, & Chen, 2008; Diamond, 2013; Shallice & Burgess, 1991). Hence, reliable assessment of EF is of great importance in clinical neuropsychology. Despite this, the psychometric properties of EF tasks are often unsatisfactory (Rabbit, 1997; Strauss, Sherman, & Spreen, 2006). To some extent, this holds also true for the TOL that is administered in a multitude of variants and versions (Berg & Byrd, 2002). For instance, for the original 12-item TOL problem set, already Humes, Welsh, and Retzlaff (1997) reported
a Cronbach’s $\alpha = 0.25$ and split-half reliability of $r = .19$, whereas Kafer and Hunter (1997) have generally questioned its validity given that a latent one-factor model of planning ability based on several TOL variables could not be fitted. Notably, this original 12-item TOL problem set is implemented in the Stockings of Cambridge (SOC) variant of the TOL in the CANTAB (Cambridge Neuropsychological Test Automated Battery; e.g., Owen, Downes, Sahakian, Polkey, & Robbins, 1990; Owen et al., 1995; Robbins et al., 1998), which is one of the most widely used, computer-assisted neuropsychological test batteries for the assessment of cognitive (dys)function in a broad range of neurological and psychiatric populations. In this respect, Lowe and Rabbitt (1998) reported test–retest reliabilities between $r = .26$ and $.60$ for different SOC scores in healthy older adults, which concurs with recent findings of Syväoja and colleagues (2015) on a test–retest reliability of $r = .23$ for SOC accuracy in healthy children.

In strong contrast to the widespread popularity of the TOL and its variants, reflected by the constantly rising number of currently almost 500 PubMed-listed publications (cf. Kaller, Rahm, Köttering, & Unterrainer, 2011) (A literature search in PubMed [http://www.ncbi.nlm.nih.gov/pubmed/] using the key term “[Tower of London] OR [Stockings of Cambridge]” revealed 494 publications [November 10, 2015]), only a remarkably small number of studies have addressed the development and evaluation of alternative problem sets in order to improve the task’s psychometric properties (e.g., Culbertson & Zillmer, 1998a, 1998b; Kaller, Unterrainer, & Stahl, 2012; Köstering, Nitschke, Schumacher, Weiller, & Kaller, 2015; Schnirman, Welsh, & Retzlaff, 1998; Tunstall, O’Gorman, & Shum, 2014). A common approach to increase task reliability is to select items from a larger pool based on the item-total correlations and to reevaluate the internal consistency of the resulting set of items in an independent sample (cf. Culbertson & Zillmer, 1998a, 1998b; Schnirman et al., 1998; Tunstall et al., 2014). As a complementary approach, the theory-grounded compilation of a consistent item set builds on knowledge and assumptions from experimental psychology about the cognitive processes underlying successful task performance and the relevant task parameters that impose a systematic variation of item difficulty (see Primi, 2014, for an instructive example on assessing fluid intelligence).

As for the TOL, task difficulty is commonly explained by the minimum number of moves necessary to solve a given problem. Yet, numerous studies have identified several additional task parameters that considerably influence the difficulty of individual TOL problems (e.g., Berg, Byrd, McNamara, & Case, 2010; Carder, Handley, & Perfect, 2004; Kaller, Unterrainer, Rahm, & Halsband, 2004; Newman & Pittman, 2007; Unterrainer, Rahm, Halsband, & Kaller, 2005; Ward & Allport, 1997; for a review, see Kaller, Rahm, Köttering, et al., 2011). To give an illustrative example for this impact of problem structure on planning beyond the minimum number of moves, Fig. 1A and B depicts two five-move TOL problems that substantially differ in the imposed demands on planning ability and, as a direct consequence, also differ in their resulting item difficulties. Figure 1C further demonstrates that six-move TOL problems, dependent on their underlying structural properties, are not necessarily harder to solve than five-move TOL problems (cf. Kaller, Unterrainer, & Stahl, 2012).

Based on comprehensive problem space analyses and empirical data, Kaller, Rahm, Köttering, et al. (2011) have, therefore, suggested a standard TOL problem set consisting of four-, five-, six-, and seven-move TOL problems (eight problems each, 32 items in total) and comprising a systematic variation of several main structural task parameters while further parameters were kept constant. This theory-grounded compilation of a TOL problem set resulted in a satisfactory increase in the split-half reliability ($r = .72$) and internal consistency ($\alpha = 0.69$) of TOL accuracy (Kaller, Unterrainer, & Stahl, 2012). The broad range of item difficulty together with an almost perfectly linear decrease in accuracy with increasing level of minimum moves was further shown to discriminate well between high- and low-performing subjects (Kaller, Unterrainer, & Stahl, 2012). However, item-specific analyses also indicated potential for further improvement, which resulted in the development of the computerized TOL-Freiburg Version (TOL-F; Kaller, Unterrainer, Kaiser, Weisbrod, & Aschenbrenner, 2012). The TOL-F constitutes an adaptation of the standard 32-item TOL problem set suggested by Kaller, Rahm, Köstering, et al. (2011) and was specifically improved to be suitable for diagnostic purposes and clinical use. In this regard, the systematic selection of 24 instead of 32 problems for the TOL-F aimed at...
achieving a reasonable trade-off between a satisfactory reliability, a sufficiently broad range of item difficulties, and an adequate test economy in terms of a relatively short (and clinically practicable) test duration.

Here, we present psychometric data on the TOL-F from two large-scale samples, which were collected in Mainz, Germany and Vienna, Austria. Given the substantial reduction from 32 to 24 items, a main objective concerned whether planning accuracy measured with the TOL-F features a comparably adequate reliability as with its more extensive precursor (Kaller, Unterrainer, & Stahl, 2012). As previous evaluations with the standard TOL problem set were solely based on student samples with a restricted representativeness, we further aimed to investigate whether the TOL-F’s psychometric properties generalize to the population. More specifically, we assessed whether the TOL-F ensures a reliable assessment of individual planning ability irrespective of a subject’s age. Finally, although previous reports have revealed substantial differences in planning ability across the life span (e.g., Albert & Steinberg, 2011; Allamanno, Della Sala, Laiacona, Pasetti, & Spinnler, 1987; Bugg, Zook, DeLosh, Davalos, & Davis, 2006; De Luca et al., 2003; Köstering, Stahl, Leonhart, Weiller, & Kaller, 2014; Krikorian, Bartok, & Gay, 1994; Luciana, Collins, Olson, & Schissel, 2009; Pen˜a-Casanova et al., 2009; Robbins et al., 1998; Zook, Welsh, & Ewing, 2006), results on sex differences are equivocal: Some studies report significant effects (De Luca et al., 2003; Peña-Casanova et al., 2009; Rönnlund, Lövden, & Nilsson, 2001; Unterrainer et al., 2013), whereas most studies have failed to find performance differences between male and female children and adolescents (Albert & Steinberg, 2011; Culbertson & Zillmer, 1998b; Krikorian et al., 1994; Luciana et al., 2009; Tunstall et al., 2014) or adults (Allamanno et al., 1987; Tunstall et al., 2014; Zook et al., 2006). Thus, another intention was to investigate sex-related differences in planning ability across life span.

Methods

Samples

Mainz sample. The Mainz sample originated from assessments with the TOL-F that were part of a large-scale prospective and population-based epidemiological study conducted at the University Medical Center Mainz with primary focus on cardiovascular health (Gutenberg Health Study; http://www.gutenberghealthstudy.org/). In the baseline examination between April 2007 and March 2012, the Gutenberg Health Study assessed a representative population sample of approximately 15,000 individuals from the city of Mainz and the district of Mainz-Bingen, Germany, which was equally stratified for sex, residence (urban and rural), and age decades. These individuals are currently being reassessed in a second examination that started in April 2012 and for which the TOL-F version as described subsequently was added in June 2012 as a measure of complex cognition. The present analyses were based on a total of 3,804 subjects who participated in the second run of the Gutenberg Health Study and were tested between June 14, 2012 and December 19, 2013. Subjects’ age ranged between 40 and 80 years. Exclusion criteria concerned insufficient knowledge of the German language and physical or psychological inability to participate in the examinations at the study center. The Gutenberg Health Study was approved by local ethics authorities. Written informed consent was obtained from all subjects prior to participation.

Data inspection revealed 34 cases (0.894%) of the Mainz sample with a null performance, indicating that no meaningful data had been collected. These were excluded before the analyses so that the final Mainz sample consisted of \( n = 3,770 \) included subjects.

Vienna sample. The Vienna sample was collected by the SCHUHFRIED GmbH as part of various norm and validation studies. The present analyses were based on a total of 830 subjects that were tested between May 30, 2011 and June 3, 2013. Subjects’ age ranged between 16 and 84 years. Participants were mainly recruited from the subject database of the SCHUHFRIED Test and Research Center constituting a nonprobability selection. Exclusion criteria concerned reports of severe neurological or psychiatric diseases and/or insufficient familiarity with using a computer mouse (as assessed with a specialized computerized test before the testing was started). Data acquisition complied with local institutional research standards for human research and was completed in accordance with the Helsinki Declaration (http://www.wma.net/en/30publications/10policies/b3/). Written informed consent was obtained from all subjects prior to participation.

Age groups and sample characteristics. Besides individual age and sex, subjects were also characterized by their highest achieved education level assessed by a five-item scale with the following levels: An educational level of 1 corresponded to 8 or less years of schooling and was typically applied to participants who completed elementary school, but did not obtain higher education. An educational level of 2 was used to classify participants who completed 9 years of schooling, but without vocational training. An educational level of 3 corresponded to 10–12 years of education and the completion of vocational training. An educational level of 4 was used to denote the completion of high school and the qualification for university entrance. An educational level
of 5 was assigned if a participant had obtained an academic degree. Information on education level was available for all but one subject from the Mainz sample.

In order to assess effects of age on planning ability, both samples were further subdivided into age groups. The Mainz sample \((n = 3,770)\) was divided into eight 5-year groups between 40 and 80 years of age, covering an age range from mid- to late adulthood. Given the smaller overall sample size of the Vienna sample \((n = 830)\), the respective age cohorts were formed in 10-year intervals ranging from 20 to 70 years of age. Subjects in the Vienna sample with an age below 20 and above 70 years comprised two additional age groups that had however substantially smaller sizes than the other age groups and that did not exactly cover the intended 10-year intervals (ranging from 16.00 to 19.83 and 70.17 to 84.00 years of age, respectively). An overview on the descriptive information for age, sex, and education level of the two overall samples as well as of the resulting subgroups is provided in Table 1.

### Tower of London-Freiburg Version

**Task description.** The TOL-F (Kaller, Unterrainer, Kaiser, et al., 2012) is implemented in the Vienna Test System (http://www.schuhfried.com/vienna-test-system-vts/) as a computerized pseudo-realistic representation of the originally wooden configuration of the TOL (Fig. 2) as it was introduced by Shallice (1982). That is, the tower configurations in the TOL-F consist of three rods of different height with the left, middle, and right rod being capable to accommodate up to three, two, and one ball, respectively. The tower configurations further comprise three balls colored in red, yellow, and blue. Thus, the green ball from the original TOL version was replaced by a yellow ball in the TOL-F in order to ensure applicability to subjects with red-green dyschromatopsia.

As in most disc-transfer planning tasks, individual problem items consist of a start and a goal state. In the TOL-F, these are presented in the lower and upper halves of the computer screen, respectively. Subjects have to transform the start into the goal state in the minimum number of moves which, in the TOL-F, is always indicated to the left of the start state. Written instructions indicate 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. Note that 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.

In order to transfer the start into the goal state, the TOL-F can be worked on by computer mouse (Vienna sample) or by touch screen (Mainz sample). In these two possible response modes, a ball is picked up either by placing the cursor over the ball and

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sample descriptives</th>
<th>Reliability estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n)</td>
<td>Age (M \pm SD) (years)</td>
</tr>
<tr>
<td>Mainz Overall sample</td>
<td>3,770</td>
<td>59.96 \pm 10.66</td>
</tr>
<tr>
<td>40.00–44.99 years</td>
<td>350</td>
<td>43.01 \pm 1.25</td>
</tr>
<tr>
<td>45.00–49.99 years</td>
<td>466</td>
<td>47.50 \pm 1.49</td>
</tr>
<tr>
<td>50.00–54.99 years</td>
<td>539</td>
<td>52.35 \pm 1.50</td>
</tr>
<tr>
<td>55.00–59.99 years</td>
<td>556</td>
<td>57.43 \pm 1.49</td>
</tr>
<tr>
<td>60.00–64.99 years</td>
<td>525</td>
<td>62.47 \pm 1.34</td>
</tr>
<tr>
<td>65.00–69.99 years</td>
<td>490</td>
<td>67.43 \pm 1.40</td>
</tr>
<tr>
<td>70.00–74.99 years</td>
<td>493</td>
<td>72.57 \pm 1.32</td>
</tr>
<tr>
<td>75.00–79.99 years</td>
<td>351</td>
<td>77.16 \pm 1.39</td>
</tr>
<tr>
<td>Vienna Overall sample</td>
<td>830</td>
<td>43.46 \pm 16.91</td>
</tr>
<tr>
<td>Below 20.00 years</td>
<td>67</td>
<td>17.79 \pm 1.15</td>
</tr>
<tr>
<td>20.00–29.99 years</td>
<td>148</td>
<td>24.94 \pm 2.70</td>
</tr>
<tr>
<td>30.00–39.99 years</td>
<td>158</td>
<td>34.50 \pm 3.18</td>
</tr>
<tr>
<td>40.00–49.99 years</td>
<td>147</td>
<td>44.96 \pm 2.75</td>
</tr>
<tr>
<td>50.00–59.99 years</td>
<td>136</td>
<td>54.57 \pm 2.71</td>
</tr>
<tr>
<td>60.00–69.99 years</td>
<td>117</td>
<td>64.60 \pm 2.73</td>
</tr>
<tr>
<td>Above 70.00 years</td>
<td>57</td>
<td>74.14 \pm 3.42</td>
</tr>
</tbody>
</table>

*Note: Underscores with bold values and solid lines denote the indices with highest and lowest estimates on reliability for the respective (sub)samples. Ed. Lvl., education level.*
clicking the left button of the computer mouse or simply by clicking the ball via finger touch. The selected ball is then encircled by a transparent whitish corona and can be moved to another rod. The respective rod is likewise selected either by computer mouse or by finger touch.

Three different time limits per trial can be applied (1 min, 3 min, unlimited) in order to accommodate differences in cognitive and/or motor speed. For the present investigation, the time limit was set to 1 min, which corresponds to the original suggestion of Shallice (1982) and was found to be sufficient, for instance, for assessing planning ability in healthy children (Unterrainer et al., 2015), young (Kaller, Unterrainer, & Stahl, 2012) and elderly healthy adults (Köstering, Stahl, et al., 2014), child patients with autism spectrum disorders and attention deficit hyperactivity disorder (Unterrainer et al., in press), Parkinson patients (McKinlay et al., 2008), as well as patients suffering from stroke and mild cognitive impairment (Köstering, Schmidt, et al., 2015). In order to avoid unnecessary frustration (and a reduced compliance and/or motivation in subsequent tests, for instance, in a clinical setting), the TOL-F allows for an automatic cancellation of the test if the time limit is exceeded three times in a row. Given a linear increase of item difficulty across the minimum number of moves (cf. Kaller, Unterrainer, & Stahl, 2012), the rationale is that subjects who struggle to cope with a certain level of item difficulty will also do so in even more complex problems. Automatic cancellation was activated for the assessments both in the Mainz and Vienna samples. Further note that for the TOL-F assessments in the Mainz sample, a time limit of 20 min for the test administration (exclusive of instructions and pauses between problem items) was requested by the study board so as to avoid delays in the subjects’ schedules (see also Experimental Procedures).

Problem set. The selection of problem items for the TOL-F was based on psychometric evaluations (Kaller, Unterrainer, & Stahl, 2012) of a recently suggested standard problem set of four- to seven-move problems for investigations with the TOL task (Kaller, Rahm, Köstering, et al., 2011). The basic idea behind the suggested composition was to develop a problem set that (i) considerably enhances the poor psychometric properties of the original TOL problem set (Humes et al., 1997; see also Kafer & Hunter, 1997; Lowe & Rabbitt, 1998; Syväöja et al., 2015) and that (ii) yields a broad range of problem difficulty—but with a linear increase across the minimum number of moves—to allow for diagnostic and research applications in a multitude of healthy and clinical samples with inherently different levels of planning ability. To this end, the structural problem parameters search depth and goal hierarchy accounting for most variation in problem difficulty were systematically varied within the levels of minimum moves, whereas other problem parameters were kept as constant as possible (for more details, see Kaller, Rahm, Köstering, et al., 2011; Kaller, Unterrainer, & Stahl, 2012). As intended, the standard problem set was found to have a satisfactory reliability (split-half $r = .72$; Cronbach $\alpha = 0.69$) and to follow a perfectly linear increase of difficulty. However, item-specific analyses indicated several possibilities for further improvement. In consequence, one of the nested families of problem items (see Kaller, Unterrainer, & Stahl, 2012) was exchanged during the development of the TOL-F, the seven-move problems were removed due to their minor contributions for discriminating high- and low-achieving subjects, and the presentation order of problems within each level of minimum moves was altered. The TOL-F problem set hence constitutes a refinement of this previously suggested standard problem set and consists of four-, five-, and six-move problems (eight problem items each). The final selection of problem items is listed in Table 2.

Before examination of planning ability using these problems, task instructions are explained and verified in the TOL-F in a prelude with simple two- and three-move problems. Planning ability can be assessed with three parallel forms, with problems across parallel forms being structurally identical, but featuring systematic permutations of ball colors, hence resulting in visually dissimilar problems that are, however, completely equivalent in terms of problem difficulty (cf. Berg & Byrd, 2002; Unterrainer
et al., 2005). That is, although all participants of the Mainz and Vienna samples were tested with the parallel form A, derived psychometric properties can be nonetheless applied to parallel forms B and C.

**Experimental procedures.** Participants of the Mainz sample were tested individually in a quiet air-conditioned room at the Gutenberg Health Study Center of the University Medical Center Mainz. Each participant was instructed by an experienced examiner who was present during the whole testing session. To overcome resistance against computer usage and to facilitate first-time handling of the computer especially in older and inexperienced participants, the Mainz sample was assessed by using a touch-sensitive screen as input device. The duration of the overall test session was limited to 20 min (exclusive of instructions and pauses between problem items) and a time limit of 60 s was specified for completion of each single problem.

The data collection for the Vienna sample was carried out as part of standardization studies at the SCHUHFRIED Test and Research Center and at the Teaching & Research Lab of the University of Vienna. All data were collected in supervised testing sessions with small groups of two to five subjects at maximum, which typically took 60–90 min to complete (inclusive additional tests). Data for the TOL-F were collected with a test version that used a computer mouse as input device. This test version did not contain an overall time limit, but a time limit of 60 s was imposed for completing each trial.

Taken together, data collections in Mainz and Vienna differed in the following three aspects: Individual testing versus small groups, touch screen versus computer mouse as input device, and the application of a 20 min overall time limit versus no such restrictions, respectively.

**Dependent measures.** For assessment of individual planning ability with the TOL-F, overall planning accuracy, defined as the percentage of problems that were correctly solved in the minimum number of moves, is regarded as the primary outcome variable of interest. It is one of the most commonly used performance measures in disc-transfer planning tasks such as the TOL (Sullivan, Riccio, & Castillo, 2009; see also Berg & Byrd, 2002) and the one yielding the greatest effect size in a meta-analysis on different outcome measures (Sullivan et al., 2009). In addition to overall planning accuracy, the TOL-F provides also information on planning accuracy separately for the different levels of minimum moves which was used for the present analyses of variance (ANOVAs) of the effects of age, sex, and task difficulty (in terms of the minimum number of moves) on planning ability (see Analyses of Variance).

Table 2. Item characteristics of the standardized TOL-F problem set

<table>
<thead>
<tr>
<th>Item</th>
<th>Start state</th>
<th>Goal state</th>
<th>Minimum moves</th>
<th>Search depth</th>
<th>Goal hierarchy</th>
</tr>
</thead>
<tbody>
<tr>
<td>#01</td>
<td>54</td>
<td>31</td>
<td>4</td>
<td>1 (high)</td>
<td>Unambiguous</td>
</tr>
<tr>
<td>#02</td>
<td>42</td>
<td>23</td>
<td>4</td>
<td>0 (low)</td>
<td>Partially ambiguous</td>
</tr>
<tr>
<td>#03</td>
<td>34</td>
<td>13</td>
<td>4</td>
<td>1 (high)</td>
<td>Partially ambiguous</td>
</tr>
<tr>
<td>#04</td>
<td>12</td>
<td>65</td>
<td>4</td>
<td>0 (low)</td>
<td>Completely ambiguous</td>
</tr>
<tr>
<td>#05</td>
<td>55</td>
<td>41</td>
<td>4</td>
<td>1 (high)</td>
<td>Unambiguous</td>
</tr>
<tr>
<td>#06</td>
<td>16</td>
<td>24</td>
<td>4</td>
<td>0 (low)</td>
<td>Partially ambiguous</td>
</tr>
<tr>
<td>#07</td>
<td>25</td>
<td>32</td>
<td>4</td>
<td>1 (high)</td>
<td>Partially ambiguous</td>
</tr>
<tr>
<td>#08</td>
<td>36</td>
<td>15</td>
<td>4</td>
<td>0 (low)</td>
<td>Completely ambiguous</td>
</tr>
<tr>
<td>#09</td>
<td>33</td>
<td>11</td>
<td>5</td>
<td>2 (high)</td>
<td>Unambiguous</td>
</tr>
<tr>
<td>#10</td>
<td>53</td>
<td>13</td>
<td>5</td>
<td>1 (low)</td>
<td>Partially ambiguous</td>
</tr>
<tr>
<td>#11</td>
<td>23</td>
<td>43</td>
<td>5</td>
<td>2 (high)</td>
<td>Partially ambiguous</td>
</tr>
<tr>
<td>#12</td>
<td>55</td>
<td>15</td>
<td>5</td>
<td>1 (low)</td>
<td>Completely ambiguous</td>
</tr>
<tr>
<td>#13</td>
<td>42</td>
<td>21</td>
<td>5</td>
<td>2 (high)</td>
<td>Unambiguous</td>
</tr>
<tr>
<td>#14</td>
<td>54</td>
<td>34</td>
<td>5</td>
<td>1 (low)</td>
<td>Partially ambiguous</td>
</tr>
<tr>
<td>#15</td>
<td>52</td>
<td>35</td>
<td>5</td>
<td>2 (high)</td>
<td>Partially ambiguous</td>
</tr>
<tr>
<td>#16</td>
<td>55</td>
<td>35</td>
<td>5</td>
<td>1 (low)</td>
<td>Completely ambiguous</td>
</tr>
<tr>
<td>#17</td>
<td>22</td>
<td>41</td>
<td>6</td>
<td>3 (high)</td>
<td>Unambiguous</td>
</tr>
<tr>
<td>#18</td>
<td>34</td>
<td>53</td>
<td>6</td>
<td>2 (low)</td>
<td>Partially ambiguous</td>
</tr>
<tr>
<td>#19</td>
<td>42</td>
<td>63</td>
<td>6</td>
<td>3 (high)</td>
<td>Partially ambiguous</td>
</tr>
<tr>
<td>#20</td>
<td>46</td>
<td>25</td>
<td>6</td>
<td>2 (low)</td>
<td>Completely ambiguous</td>
</tr>
<tr>
<td>#21</td>
<td>23</td>
<td>61</td>
<td>6</td>
<td>3 (high)</td>
<td>Unambiguous</td>
</tr>
<tr>
<td>#22</td>
<td>43</td>
<td>64</td>
<td>6</td>
<td>2 (low)</td>
<td>Partially ambiguous</td>
</tr>
<tr>
<td>#23</td>
<td>13</td>
<td>32</td>
<td>6</td>
<td>3 (high)</td>
<td>Partially ambiguous</td>
</tr>
<tr>
<td>#24</td>
<td>22</td>
<td>55</td>
<td>6</td>
<td>2 (low)</td>
<td>Completely ambiguous</td>
</tr>
</tbody>
</table>

**Note:** Start and goal states are reported in the notation suggested by Berg and Byrd (2002). For search depth, numbers indicate the number of initially to be accomplished intermediate moves, whereas descriptions in parentheses refer to the level of search depth with regard to the respective minimum number of moves. For further information on structural problem parameters in the TOL task, please refer to the overview provided in Kaller, Rahm, Köstering, et al. (2011).
Secondary variables of the TOL–F concern the total number of problems solved at all (irrespective of whether they are solved in the minimum number), and the initial thinking and movement execution times, which will be reported and discussed elsewhere. Further note that the TOL–F also provides reports on the number and type of observed rule breaks (e.g., trying to select a blocked ball, to place a ball on a blocked rod or outside the tower configuration) which may be informative for clinical use so as to disentangle a patient’s impairments in planning ability. However, given their infrequent occurrence in healthy subjects, these data were not subject to the present data analyses.

Data Analyses

Analyses of variance. Mixed ANOVAs on planning accuracy as dependent variable were conducted using IBM SPSS Statistics for Windows (Version 21.0; IBM Corp., Armonk, NY) to test for main effects of the between-subjects factors “Age Groups” and “Sex” and the within-subject factor “Minimum Moves” and their interactions. ANOVAs were separately run for the Mainz and Vienna samples.

Reliability estimates. In accordance with the revised review model for the description and evaluation of psychological and educational tests (Version 4.2.6; http://www.efpa.eu/professional-development/assessment) recently suggested by the Board of Assessments of the European Federation of Psychologists’ Associations (EFPA, 2013), the following estimates of reliability were considered for assessing the internal consistency of overall planning accuracy as the primary TOL–F outcome variable: \( \lambda_2 \), \( \lambda_3 \) reflecting Cronbach \( \alpha \), \( \omega_{tot} \), and the greatest lower bound (glb). In addition, \( \lambda_4 \) was computed for comparability with the exhaustive split-half reliability estimates reported by Kaller, Unterrainer, and Stahl (2012). All indices were computed using the psych package (Version 1.3.2; Revelle, 2013) for the R open-source statistical software (Version 3.0.1; R Core Team, 2013). Note that although most commonly reported, the \( \lambda_3 \) measure (or Cronbach \( \alpha \)) is only a lower bound to the reliability that often constitutes a gross underestimate (Cortina, 1993; Sijtsma, 2009). Sijtsma (2009), therefore, advocated using glb as a better alternative (Bentler & Woodward, 1980), whereas Revelle and Zinbarg (2009) put forward \( \omega_{tot} \) (McDonald, 1999).

Reliability indices were separately computed for the two overall samples from Mainz and Vienna as well as for the respective age subgroups (cf. Age Groups and Sample Characteristics). Given that adequate reliability estimations require sufficiently large sample sizes of at least 100 and preferably more than 200 subjects per (sub)group (EFPA, 2013), a further subdivision along sex and education level was omitted. In this respect, even the estimates for the different Vienna age subgroups have to be interpreted with some caution. However, as data collection is continuing for the Mainz sample, comprehensive reliability estimates for age, sex, and education-fair subsamples will be provided once the necessary numbers of observations are available.

Results

Effects of Age, Sex, and Minimum Moves on Planning Accuracy

Mainz sample. A mixed ANOVA with the between-subjects factors Age Groups (cf. Age Groups and Sample Characteristics) and Sex (men vs. women) and the within-subject factor Minimum Moves (four-, five-, and six-move problems) on planning accuracy as dependent variable revealed significant main effects of Age Groups \( (F_{(7,3754)} = 115.44, p < .001, \eta^2_{partial} = 0.177) \), Sex \( (F_{(1,3754)} = 114.31, p < .001, \eta^2_{partial} = 0.030) \), and Minimum Moves \( (F_{(2,7508)} = 10610.58, p < .001, \eta^2_{partial} = 0.739) \). Also the two-way interactions of Age Groups by Minimum Moves \( (F_{(14,7508)} = 12.33, p < .001, \eta^2_{partial} = 0.022) \) and Sex by Minimum Moves \( (F_{(2,6034)} = 15.62, p < .001, \eta^2_{partial} = 0.004) \) reached significance, whereas Age Groups by Sex \( (F_{(7,3754)} = 0.67, p = .702, \eta^2_{partial} = 0.001) \) and the three-way interaction were not found to be significant \( (F_{(14,7508)} = 0.62, p = .847, \eta^2_{partial} = 0.001) \). Linear contrasts further confirmed that planning accuracy systematically declined in a linear fashion both with increasing age \( (F_{(1,3754)} = 736.77, p < .001, \eta^2_{partial} = 0.164) \) and with increasing task complexity in terms of the minimum number of moves for optimal solution \( (F_{(1,3754)} = 21383.62, p < .001, \eta^2_{partial} = 0.851) \). Linear contrasts also revealed that the factor Minimum Moves exerted a linear moderator effect on both the main effects of Age Groups \( (F_{(7,3754)} = 24.74, p < .001, \eta^2_{partial} = 0.044) \) and of Sex \( (F_{(1,3754)} = 31.45, p < .001, \eta^2_{partial} = 0.008) \). That is, effects of age and sex on planning accuracy linearly increased with increasing TOL–F task difficulty and were hence more pronounced in the six-move problems.

The main results in the Mainz sample are illustrated in Fig. 3A (left panel, upper right panel), which also shows that overall planning accuracy approached a normal distribution (Fig. 3A, upper middle panel; \( M \pm SD, 55.93 \pm 15.60\% \)) with both “skewness” \( (−0.208; SE, 0.040) \) and “kurtosis” \( (−0.211; SE, 0.080) \) approximating a value of zero. Moreover, the intended linear increase of task difficulty in the TOL–F was closely reflected by the individual limits of subjects’ planning performance (Fig. 3A, lower right panel). Thus, subjects with a low, medium, and high level of planning ability were likely to attain correct solutions in problems up to a level of four, five, and six minimum moves, respectively, but not above. In other words, this important
aspect of construct validity of the theory-driven TOL-F problem selection was demonstrated by the observed empirical task difficulty of the minimum move subscales.

Vienna sample. The same mixed ANOVA as earlier was also conducted for the Vienna sample revealing significant main effects of Age Groups ($F_{6,816} = 6.08, p < .001, \eta^2_{\text{partial}} = 0.043$), Sex ($F_{1,816} = 6.37, p = .012, \eta^2_{\text{partial}} = 0.008$), and Minimum Moves ($F_{2,1632} = 1837.20, p < .001, \eta^2_{\text{partial}} = 0.692$) as well as a significant two-way interaction of Age Groups by Minimum Moves ($F_{12,1632} = 1.98, p = .012, \eta^2_{\text{partial}} = 0.014$). The remaining two-way interactions of Age Groups by Sex
The main objective of the present study was to evaluate the reliability of assessing planning ability across the lifespan with the TOL (Freiburg version; TOL-F), whereas the second objective was to assess age- and sex-related differences in planning ability. With regard to the first objective, reliable assessment of EF and related complex cognitive abilities is often difficult to achieve due to the multifaceted nature of EF, their inherent task impurity, and reliance on novel situations that elicit nonroutine behavior (Rabbitt, 1997; Strauss et al., 2006). Therefore, EF tasks need to be sufficiently complex—while at the same time covering a broad range of problem-related processing time for all subjects. However, for evaluating the TOL-F’s practicability in clinical contexts, the total test durations inclusive of instructions and pauses are more informative and relevant and, in consequence, are descriptively provided in Fig. 4 for the overall samples from Mainz and Vienna as well as for the respective age cohorts. Overall medians (Md) and distributions (percentile ranks, PR) of the total durations were highly comparable in both samples with TOL-F test administrations lasting between 13 min (PR 15) and 20 min (PR 85) in the majority of the Mainz subjects (Md = 17 min) and between 12 and 20 min in the majority of the Vienna subjects (Md = 16 min). Furthermore, as becomes obvious from Fig. 4, total test durations of TOL-F administrations increase with older age.

Likewise, the percentage of test cancellations due to the applied criteria of three subsequent timeouts (both samples) and the 20-min overall time limit (Mainz sample only) rose with increasing age, particularly from 40 years onwards (Fig. 5). Interestingly, the pattern for the three subsequent timeouts limit appears to mark two stages (around problem items #13 and #20) with an increased frequency of test cancellations with older age, which may reflect declining capacities for efficiently planning ahead with increasing age (Fig. 5; see also Discussion).

Discussion

The main objective of the present study was to evaluate the reliability of assessing planning ability across the lifespan with the TOL (Freiburg version; TOL-F), whereas the second objective was to assess age- and sex-related differences in planning ability. With regard to the first objective, reliable assessment of EF and related complex cognitive abilities is often difficult to achieve due to the multifaceted nature of EF, their inherent task impurity, and reliance on novel situations that elicit nonroutine behavior (Rabbitt, 1997; Strauss et al., 2006). Therefore, EF tasks need to be sufficiently complex—while at the same time covering a broad range of difficulty—and to comprise a relatively large number of items to yield valid and reliable estimates at the first administration. This, however, is often not achievable in clinical routine, where easy-to-administer, short tests are required. As a result, extant TOL versions for use in clinical settings have mostly yielded insufficient or inconsistent reliability estimates (e.g., Humes et al., 1997; Lowe & Rabbitt, 1998; Syväoja et al., 2015; but see Culbertson & Zillmer, 1998a, 1998b) and provoked criticism of the validity of the TOL (Kafer & Hunter, 1997). In contrast, here it was demonstrated that split-half reliability and internal consistency of TOL-F accuracy was adequate overall as well as across a wide age range, as the different reliability estimates ranged between 0.713 and 0.755 for the overall Mainz sample and were only slightly lower (~0.05 units) for the overall Vienna sample (0.656–0.730). Moreover, despite differences in test administration (touch screen vs. computer mouse; overall time limit of 20 min vs. no time limit; see Methods), reliability estimates were highly comparable between the two samples, for instance, the glb converged to

$F_{(6,816)} = 1.45, p = .194, \eta^2_{partial} = 0.011$) and Sex by Minimum Moves ($F_{(2,1632)} = 1.18, p = .309, \eta^2_{partial} = 0.001$) and the three-way interaction ($F_{(12,1632)} = 0.87, p = .581, \eta^2_{partial} = 0.006$) failed to reach significance. Linear contrasts again confirmed that planning accuracy linearly decreased with increasing age ($F_{(1,816)} = 25.92, p < .001, \eta^2_{partial} = 0.031$) and increasing task difficulty in terms of minimum number of moves to solution ($F_{(1,816)} = 3422.52, p < .001, \eta^2_{partial} = 0.807$) and that the former effect was linearly moderated by the latter ($F_{(6,816)} = 2.49, p = .021, \eta^2_{partial} = 0.018$). The main results in the Vienna sample are illustrated in Fig. 3B (left panel, upper right panel) and complemented by showing that overall planning accuracy also followed a normal distribution for these data (Fig. 3B, upper middle panel; $M \pm SD = 61.25\% \pm 14.31\%$) with both skewness ($-0.303; SE, 0.085$) and kurtosis ($-0.086; SE, 0.170$) approaching a value of zero. Furthermore, task difficulty in the TOL-F and the assessed planning ability at the level of individual subjects were closely linked also for the Vienna sample (Fig. 3B, lower right panel), thus again suggesting construct validity of the TOL-F.

Reliability Estimates for Overall Planning Accuracy

A comprehensive overview on the reliability estimates is provided in Table 1. The five different estimates on reliability ranged between 0.713 and 0.755 for the overall Mainz sample, but were slightly lower (~0.05 units) for the overall Vienna sample (0.656–0.730). In both the overall samples and in the respective age groups, estimates were highest for glb and $\lambda_4$, whereas $\lambda_3$ or Cronbach’s $\alpha$ yielded the lowest estimate in all cases. Notably, glb converged for the Mainz and the Vienna overall samples to almost congruent estimates of 0.755 and 0.730, respectively. Taken together, these results suggest that the TOL-F features an adequate and satisfactory reliability with estimates based on glb and $\lambda_4$ always exceeding 0.7 for the two samples as well as for all age-related subsamples (Table 1).

Total Test Durations and Test Cancellations

The 20-min time limit for the TOL-F assessments in the Mainz Sample (see Task Description and Experimental Procedures) was exclusive of the instructions and self-determined pauses between problem items so as to ensure a comparable overall problem-related processing time for all subjects. However, for evaluating the TOL-F’s practicability in clinical contexts, the total test durations inclusive of instructions and pauses are more informative and relevant and, in consequence, are descriptively provided in Fig. 4 for the overall samples from Mainz and Vienna as well as for the respective age cohorts. Overall medians (Md) and distributions (percentile ranks, PR) of the total durations were highly comparable in both samples with TOL-F test administrations lasting between 13 min (PR 15) and 20 min (PR 85) in the majority of the Mainz subjects (Md = 17 min) and between 12 and 20 min in the majority of the Vienna subjects (Md = 16 min). Furthermore, as becomes obvious from Fig. 4, total test durations of TOL-F administrations increase with older age.

Likewise, the percentage of test cancellations due to the applied criteria of three subsequent timeouts (both samples) and the 20-min overall time limit (Mainz sample only) rose with increasing age, particularly from 40 years onwards (Fig. 5). Interestingly, the pattern for the three subsequent timeouts limit appears to mark two stages (around problem items #13 and #20) with an increased frequency of test cancellations with older age, which may reflect declining capacities for efficiently planning ahead with increasing age (Fig. 5; see also Discussion).
almost identical estimates of 0.755 and 0.730 for the overall Mainz and Vienna samples, respectively. Furthermore, reliability was found to be stable across different age groups, with glb and \( l_4 \) exceeding 0.7 for all subgroups. In contrast, Cronbach’s \( \alpha \) (Cronbach, 1951) yielded the lowest estimate in all cases. This is in line with suggestions that—although still the most commonly reported reliability metric—Cronbach’s \( \alpha \) may not be the most appropriate measure and will underestimate the true reliability in many cases and that \( v_{tot} \) and glb be used as better indices of the lower bound method of quantifying reliability (Peters, 2014; Revelle & Zinbarg, 2009; Sijtsma, 2009). For reasons of comparability, we nonetheless also report Cronbach’s \( \alpha \), which reached 0.713 at least in the Mainz sample and thus indicated sufficiently high reliability of the item set, thus clearly exceeding previously reported data from nonoptimized problem sets (see Humes et al., 1997; see also Kafer & Hunter, 1997). Moreover, current reliability estimates also exceed those of a recent empirically derived four-disk TOL version, which yielded maximal values of \( v_{tot} = 0.64 \), glb = 0.67, and \( l_4 \) = 0.65 in children, but lower estimates in adults (Tunstall et al., 2014). That is, using indices concuring with newly developed research criteria (EFPA, 2013), reliability estimates of accuracy of the TOL-F problem set originally composed on the basis of theoretical problem space analyses were found to be satisfactory and considerably higher than those for TOL versions based on empirical item selection. The TOL-F can, therefore, be regarded as a planning test suitable for adults, to which the four-disk TOL version by Tunstall and colleagues (2014) may represent a valuable complementary test for children.

Furthermore, the TOL-F yielded a broad range of performance even in healthy participants, with mean accuracy ranging approximately between 40% and 70% (see Fig. 3A and B, left panels). More specifically, five- and six-move problems contributed most to performance differences of the adults tested here, whereas four-move problems did not provide much variance. In clinical populations, however, it can be expected that four-move problems will account for additional variance, yielding even broader ranges of performance than found in the present sample (cf. Köstering, Schmidt, et al., 2015). As originally envisaged, the TOL-F was further shown to be a “graded difficulty test” (Shallice, 1982, p. 204), where low-performing subjects gradually fail to solve problems at more demanding levels, whereas high-performing ones reliably solve easier as well as more difficult problems (Fig. 3A and B, lower right panels), which also attests to the TOL-F’s construct validity. In close relation, recent analyses of the TOL-F problem set using the framework of Item Response Theory (cf. Hambleton & Swaminathan, 1985) underlined its construct validity for measuring planning ability in terms of a psychometrically unidimensional trait (Debelak, Egle, Köstering, & Kaller, 2015), and the test–retest reliability of TOL-F accuracy, measured as the relative consistency and absolute agreement
of individual performance over time, was also found to be adequate (Köstering, Nitschke, et al., 2015). Based on these and the present results, we propose the TOL-F as a clinically feasible, yet reliable test of complex cognition.

It has to be noted, however, that current results for the TOL-F do not necessarily generalize to other TOL versions that do not feature a systematic manipulation of structural determinants of problem difficulty. That is, as the difficulty of individual TOL problems is substantially determined by structural problem parameters beyond the minimum moves such as goal hierarchy, search depth, and the number of optimal paths to solution (e.g., Berg et al., 2010; Kaller et al., 2004; Ward & Allport, 1997; for an extensive discussion, see Kaller, Rahm, Köstering, et al., 2011), it was hence an integral part of the present theoretically grounded approach to control for these influences by keeping them at a constant level within and across the minimum number of moves (cf. Kaller, Rahm, Köstering, et al., 2011). In addition, accounting for structural problem parameters has also enabled examination
of their specific impact in neurological patients (Köstering, McKinlay, Stahl, & Kaller, 2012; McKinlay et al., 2008; Rainville, Lepage, Gauthier, Kergoat, & Belleville, 2012; see also Andrews, Halford, Chappell, Maujean, & Shum, 2014) as well as in developmental studies extending the common knowledge about planning trajectories (Kaller, Rahm, Spreer, Mader, & Unterrainer, 2008; Köstering, Stahl, et al., 2014; Unterrainer et al., 2013, 2015, in press), and also provided new insights into the neural foundation of planning ability using functional and structural brain imaging (Kaller et al., 2015; Kaller, Heinze, et al., 2012; Kaller, Rahm, Spreer, Weiller, & Unterrainer, 2011; Newman, Greco, & Lee, 2009; Ruh, Rahm, Unterrainer, Weiller, & Kaller, 2012). Thus, carefully selecting TOL items with regard to their structural problem parameters is both critical for adequate psychometric properties and useful for a more specific characterization of planning ability in healthy and clinical populations.

With respect to the second objective, this is the largest study to date to assess both age- and sex-related differences in TOL performance in healthy adults. This study is, therefore, an important complement to a recent study on TOL performance from childhood to young adult age based on n ~ 900 participants (Albert & Steinberg, 2011). More specifically, in line with previous studies (e.g., Allamanno et al., 1987; De Luca et al., 2003; Köstering, Stahl, et al., 2014; Peña-Casanova et al., 2009; Robbins et al., 1998), in the Mainz sample, there was an age-related decrease in planning ability, which unfolded in a decrement in accuracy of about 20% from age 40 to 80 (Fig. 3A). As obvious in the Vienna sample (Fig. 3B), performance remained at a stable level from ages 16 to 35 with a decrease beginning at about 45 years. Thus, taking together results from both samples, age-related performance was found to be comparably stable from late adolescence/early adulthood until the fifth decade of life and to decrease from thereon in an almost linear trajectory. Important for further diagnostic considerations, the age-related decline was strongly tied to problem difficulty, as the ability to correctly solve more difficult five- and in particular six-move problems decreased with increasing age (Fig. 3). This clearly points to the fact that in older ages, planning per se still worked for easier tasks, but when longer sequences of moves had to be generated and evaluated, limitations became obvious. Recently, it has been shown that age-related differences in TOL planning accuracy of adults aged 60 years and older are driven by concomitant differences in fluid reasoning and—to a lesser extent—in working memory capacity, but not by a general slowing in processing speed (Köstering, Leonhart, Stahl, Weiller, & Kaller, 2014). As fluid reasoning and working memory (capacity) are known to be linked to planning performance (e.g., Bugg et al., 2006; Gilhooly, Wynn, Phillips, Logie, & Della Sala, 2002; Phillips, Wynn, Gilhooly, Della Sala, & Logie, 1999; Unterrainer et al., 2004; Zook, Davalos, DeLosh, & Davis, 2004; Zook et al., 2006), whereas processing speed is an unspecified function known to affect on a range of cognitive domains (Salthouse, 1996, 2000), this further argues for an age-related deterioration in the core planning processes from mid-adulthood onward as evidenced here. Nonetheless, one caveat remains namely that—due to the cross-sectional data acquisition in both samples—the present age effects may at least to some extent be also driven by cohort and period effects and hence be to some extent overestimated.

Another issue that has to be discussed is the observed sex difference. As depicted in Fig. 3A and B, this phenomenon was observable in both samples independently and indicated that men outperformed women by about 5%. Although the sex by age interaction was not significant in either sample, in the Vienna data, it is apparent on a descriptive level that men were significantly better than women from about 45 years onward. In the Mainz sample, the sex difference was found across all age groups from 40 to 80 years. Previous reports on sex differences in planning ability using tower tasks have been very inconsistent, with few studies revealing significant effects (De Luca et al., 2003; Peña-Casanova et al., 2009; Rönnlund et al., 2001; Unterrainer et al., 2013). In contrast, present results can be taken to reveal that sex effects on planning performance do indeed exist, albeit only translating into a modest difference in accuracy. Furthermore, the fifth decade of life possibly marks a critical age during adulthood at which these sex differences begin to emerge. Interestingly, this closely parallels the trajectory of age-related decrements in planning accuracy, which also became consistently evident from 40 years of age onward. That is, reasons for overlooked sex differences in previous studies may have been due to the often limited sample size and the age of participants. In line with the latter explanation, of the four studies that did find significant differences between adult male and female participants, two studies tested samples in the age range from about 35 to 90 years (Peña-Casanova et al., 2009; Rönnlund et al., 2001), thus roughly concurring with the age range for which sex differences were evident here, whereas the remaining studies reported sex effects for the whole sample aged 8–64 years (De Luca et al., 2003) and for preschoolers (Unterrainer et al., 2013).

One explanation for the sex differences in planning accuracy might be the well-accepted common knowledge that men perform better on visuospatial tasks than women (Halpern, 1997; Kimura, 1992). However, this cannot fully account for the interaction of sex and minimum number of moves found for the Mainz sample, which revealed that differences between men and women increased as a function of problem difficulty. Although more moves and problem states have to be visualized for more complex problems, the nature of what has to be visualized does not change. Therefore, increased problem difficulty is likely to result in increased demands on working memory (more information to actively maintain and manipulate) and relational reasoning (more move interdependencies to consider), rather than in a qualitative difference in basic visuospatial processing. Thus, sex differences are more likely to arise from differences in the planning processes per se, that is, in the mental simulation and evaluation of moves and their interdependencies. In this regard, sex differences favoring men are relatively consistently reported for visuospatial working memory (e.g., Cansino et al., 2013; De Luca et al., 2003; Kaufman, 2007; Lejbak, Crossley, & Vrbancic, 2011; Pauls,
sex-related differences in planning ability that were especially pronounced from the fifth decade onward (Fig. 3) could further be
dealing with the most complex problems with six minimum moves.
increases for the older age groups, especially from 65 years of age onward. Again, as the test cancellation occurs after the first
problems (i.e., with six minimum moves), and this pattern of timeouts seems to be age related in that the frequency of subjects
mid-adulthood. The second stage where a proportion of subjects reaches three consecutive timeouts occurs at the most difficult
explained by an age-related slowing of processing speed, but is rather reflective of lower planning ability across all stages of
move and the even more difficult six-move problems, so that for them, the test cancellation spares them this frustration and
parameters.
overall limit of participants' planning ability, rather than a failure to cope with a specific (or arbitrary) combination of problem
minimum move length (and repeated in the second set of four items), subjects have worked on each possible combination of
combinations of the problem parameters search depth and goal hierarchy that are presented in these first four items of each
regard to the three consecutive timeouts, there is a marked pattern of "peaks" in the frequency at items #13 and #20 (especially
to three consecutive timeouts (in both samples) and due to the 20-min overall time limit (in the Mainz sample only). With
data acquisition with subtle time constraints that meet a realistic testing situation.
Collecting large data sets is highly desirable but almost always comes at certain trade-off, for example, group testings (Vienna) or
overall time limits (Mainz). Because all participants were tested using the identical computer program in quiet and standardized
settings, differences between group and single testing should be negligible. Furthermore, despite the predetermined overall time
limit of 20 min (exclusive of instructions and self-determined pauses between problem items) in the Mainz sample, only a small
number of participants (5.8%) did not accomplish to finalize the test within 20-min processing time (Fig. 5A), indicating that the
given time window was sufficient in the majority of cases. In comprehensive neuropsychological examinations, the time available
for single tests is typically severely restricted and thus many researchers and clinicians might even favor and benefit from a reliable
data acquisition with subtle time constraints that meet a realistic testing situation.

When inspecting the pattern of test cancellations in more detail (Fig. 5), one has to differentiate between cancellations due
to three consecutive timeouts (in both samples) and due to the 20-min overall time limit (in the Mainz sample only). With
regard to the three consecutive timeouts, there is a marked pattern of "peaks" in the frequency at items #13 and #20 (especially
in the Mainz sample; Fig. 5). These two items are the fifth and fourth item presented of the five-move and six-move problems,
respectively. That is, when the test is cancelled, subjects have either worked on the first five 5-move-items and consecutively
failed to solve them, or worked on the first four 6-move-items and consecutively failed them. Importantly, as there are four
combinations of the problem parameters search depth and goal hierarchy that are presented in these first four items of each
minimum move length (and repeated in the second set of four items), subjects have worked on each possible combination of
minimum moves, search depth, and goal hierarchy (cf. Table 2). Thus, the test cancellations are most likely the result of an
overall limit of participants' planning ability, rather than a failure to cope with a specific (or arbitrary) combination of problem
parameters.

In this regard, there seems to be a two-stage pattern of limitations in planning ability, with the first becoming obvious already at
the level of five minimum moves, whereas the second becomes obvious at six minimum moves. That is, there is a proportion of
subjects with low performance for whom the five-move problems are already a significant challenge. Without the cancellation after
three subsequent timeouts, these subjects would most probably be very frustrated at having to work through the rest of the five-
move and the even more difficult six-move problems, so that for them, the test cancellation spares them this frustration and
avoids unnecessary testing. Importantly, this pattern is not (entirely) age related, but also occurs in the "younger" groups of mid-
adolesthood from 40 years of age onward (and only begins to rise substantially after 65 years of age), so it cannot be exhaustively
explained by an age-related slowing of processing speed, but is rather reflective of lower planning ability across all stages of
mid-adulthood. The second stage where a proportion of subjects reaches three consecutive timeouts occurs at the most difficult
problems (i.e., with six minimum moves), and this pattern of timeouts seems to be age related in that the frequency of subjects
increases for the older age groups, especially from 65 years of age onward. Again, as the test cancellation occurs after the first
four items with six minimum moves have been worked on, this also seems to reflect the effect of age-related decrements in planning
ability and in dealing with the most complex problems with six minimum moves.
With regard to test cancellations due to the 20-min overall time limit in the Mainz sample, there is—by way of nature of this time limit—no clear pattern regarding the specific items at which this cancellation occurs, but there is a marked age-related increase in test cancellations. Besides reaching the individual limits in planning ability (as discussed earlier), this age-related increase may also be attributed to an unspecified decrease in processing speed with older age that is unrelated to participants’ planning ability (but see Köstering, Leonhart, et al., 2014). This might hence have led to an overestimation of the age-related decrease in planning ability in the Mainz sample (and explain the lower accuracy of the older age groups when compared with their homolog groups in the Vienna sample). But importantly, in the Vienna sample, a similar age-related decrease in planning ability was found, so that the overall time limit in the Mainz sample did not change the pattern of age-related performance differences per se (only its extent).

Finally, a recent study comparing manual and computerized administrations of the TOL (using a balanced problem set similar to the present one) did not reveal any significant differences in planning performance in healthy young adults (McKinlay & McLellan, 2011). A similar study on patients with Autism Spectrum Disorder yielded the same results, concluding that computer-administered and experimenter-administered versions can be used interchangeably (Williams & Jarrold, 2013). However, particularly for clinical applications, it might be relevant whether a manual administration effectively leads to a higher number of rule violations than a computerized administration, given that such rule violations could help to better understand and characterize a patient’s impairments in planning. Although the TOL-F does not allow breaking rules, it (i) records any attempts to do so and (ii) provides data on the number and type of observed rule breaks (see also Dependent Measures). That is, whether rule breaks of patients in manual and computerized task versions are of a different quality and/or quantity remains subject to future research, but clinically valuable information on this matter can be derived from the TOL-F.

A further, partly related question concerns the extent to which the present findings on the TOL-F’s psychometric properties generalize to clinical populations. First data on patients with ischemic stroke (n = 60), Parkinson’s syndrome (n = 51), and mild cognitive impairment (n = 29) indicate that the TOL-F constitutes a reliable and valid measure of planning ability in these groups of neurologically impaired patients (Köstering, Schmidt, et al., 2015). However, as sample sizes were small and within-sample heterogeneity can in general be expected to be enlarged in patient studies, these findings have to be consolidated by more comprehensive data collections. In this respect, psychiatric populations with well-known planning impairments such as patients with schizophrenia, depression, obsessive compulsive disorder, attention deficit hyperactivity disorder, and autism spectrum disorder should also be considered in future assessments of the TOL-F’s clinical utility.

Conclusion

Using a theoretically grounded problem set accounting for structural properties beyond the minimum number of moves, planning accuracy on the TOL-F possesses adequate psychometric properties that are stable across the adult lifespan. At the same time, TOL-F accuracy covers a broad range of graded difficulty even in healthy adults, which makes this task suitable for both research and clinical application. Furthermore, an age-related decrease in planning accuracy as well as significant differences between men and women unfold from the fifth decade onward.

Funding

The Gutenberg Health Study is funded through the government of Rhineland-Palatinate (“Stiftung Rheinland-Pfalz für Innovation,” contract AZ 961-386261/733), the research programs “Wissen schafft Zukunft” and “Center for Translational Vascular Biology (CTVB)” of the Johannes Gutenberg University of Mainz, and its contract with Boehringer Ingelheim and PHILIPS Medical Systems, including an unrestricted grant for the Gutenberg Health Study. This study was also supported by the intramural grant program “MAIFOR” of the University Medical Center of the Johannes Gutenberg University Mainz.

CPK and LK are supported by the BrainLinks-BrainTools Cluster of Excellence funded by the German Research Foundation (DFG; grant # EXC 1086). LK is further supported by scholarship funds from the State Graduate Funding Program of Baden-Württemberg, Germany. PSW is funded by the Federal Ministry of Education and Research (BMBF 01EO1003) and received honoraria for lectures or consulting from Boehringer Ingelheim and Bayer Healthcare, Leverkusen.

Conflict of Interest

CPK and JMU declare to receive a proportion of TOL-F license fees from the SCHUHFRIED GmbH due to authorship for the published TOL-F test materials (Kaller, Unterrainer, et al., 2012). RD and JE are employees of the SCHUHFRIED GmbH that maintains and commercially distributes the computerized TOL-F as part of the Vienna Test System. All other authors report no conflicts of interest.
References


