Psychological Assessment
2012, Vol. 24, No. 1, 46–53

© 2011 American Psychological Association
1040-3590/12/$12.00 DOI: 10.1037/a0025174

# Assessing Planning Ability With the Tower of London Task: Psychometric Properties of a Structurally Balanced Problem Set

Christoph P. Kaller
University Medical Center Freiburg

Josef M. Unterrainer
University Medical Center Mainz

Christoph Stahl
University of Cologne

In clinical and experimental settings, planning ability is typically assessed using the Tower of London (ToL) or one of its variants. For enhancing the comparability across studies, a common ToL problem set was recently suggested comprising a collection of 4- to 7-move problems. Based on previous theoretical and empirical analyses of problem space and task structure, development of the problem set accounted particularly for the influence of structural problem parameters on the detection of individual differences in planning ability. To assess its adequacy as a clinical and research instrument, the present study evaluated the psychometric properties of the suggested problem set. Results showed a clear and nearly perfect linear increase of task difficulty across minimum moves. Given a broad range of item difficulty, high- and low-achieving subjects could be well discriminated. The test scores' split-half reliability ($r =$ .72) and internal consistency ($\alpha = .69$) were satisfactory. Taken together, the ToL problem set evaluated here proved to have good psychometric qualities and constitutes a conceptually sound basis for diagnostic and research purposes.

*Keywords:* planning ability, Tower of London, psychometric quality, reliability, executive functions

*Supplemental materials:* http://dx.doi.org/10.1037/a0025174.supp

In many situations beyond everyday routine, successful completion of purposive behavior relies essentially on the ability to identify and select an appropriate sequence of behavior before its actual execution. Planning ahead future actions comprises the mental conception and evaluation of several behavioral alternatives and their associated consequences (Goel, 2002; Ward & Morris, 2005). It is one of the highest and most human cognitive abilities, and, as such, it depends fundamentally on the integrity of the prefrontal cortex (Owen, 2005).

In clinical and experimental neuropsychology, planning ability is assessed most often using the Tower of London (ToL) task or one of its variants (Berg & Byrd, 2002; Kaller, Rahm, Köstering, & Unterrainer, 2011). The ToL is a so called disc-transfer paradigm that was originally developed to measure planning impairments in patients with frontal lesions (Shallice, 1982). In the ToL, planning is required for an efficient transformation of a given start state into a desired goal state, that is, for an optimal solution within the minimum number of moves. The task's general scenario is knowledge-lean and well defined with explicit specification of the start state, the goal state, the transformation operators and their restrictions (Ward & Morris, 2005). The classic version of the ToL consists of three differently colored balls placed on three vertical rods of different heights that may hold at maximum either one, two or three balls, respectively (for overviews on other versions and variants, see Berg & Byrd, 2002; Hinz, Kostov, Kneißl, Sürer, & Danek, 2009).

Despite the enormous popularity of the ToL, only a few studies have systematically addressed the determinants of item difficulty that result from the structural properties of the task and its problem space (e.g., Berg, Byrd, McNamara, & Case, 2010; Carder, Handley, & Perfect, 2004; Kaller, Unterrainer, Rahm, & Halsband, 2004; Newman & Pittman, 2007; Ward & Allport, 1997). Instead, problem difficulty is usually defined only in terms of the minimum number of moves required for an efficient solution without consideration of other task factors. However, growing evidence suggests that this assumption is an inadequate approximation, given that several other structural problem parameters also exert substantial influence on the assessment of planning ability in the ToL (for a comprehensive review, see Kaller et al., 2011). That is, neither do problems with an equal minimum number of moves necessarily share identical levels of task difficulty, nor do gradual increases of minimum number of moves imply a correlated rise of

This article was published Online First August 22, 2011.

Christoph P. Kaller, Department of Neurology, University Medical Center Freiburg; Josef M. Unterrainer, Medical Psychology and Medical Sociology, University Medical Center Mainz; Christoph Stahl, Department of Psychology, University of Cologne.

This work was supported by the German Federal Ministry of Education and Research (BMBF) Grant 01GW0710. We thank Sonja Kappler for assistance in data acquisition and Nina Ruh for invaluable discussion of an earlier draft of this article.

Correspondence concerning this article should be addressed to Christoph P. Kaller, Department of Neurology, University Medical Center Freiburg, Breisacher Str. 64, DE-79106 Freiburg, Germany. E-mail: christoph.kaller@uniklinik-freiburg.de

task difficulty (cf. McKinlay et al., 2008). In contrast, problems with a higher minimum number of moves may even be considerably easier to solve than others that require less moves to solution (e.g., see Figure 1A). Item selections based on the still common one-dimensional consideration of problem difficulty in terms of minimum moves[1] may hence entail a serious paucity in the content validity of any interpretations derived from the resulting test scores given a deficient operationalization of problem difficulty. Moreover, it may also be an ultimate source of the so far only limited convergent and/or concurrent validity of using tower task performance scores within the context of diagnostic decision making (for a meta-analysis, see Sullivan, Riccio, & Castillo, 2009). In other words, the insufficient conceptualization of task demands and problem difficulty may have contributed substantially to both the inconsistencies across clinical studies (Sullivan et al., 2009) and, in this respect, to the poor psychometric properties of the ToL (e.g., Cronbach $\alpha = .25$; split-half reliability = .19; for the original ToL problem set, see Humes, Welsh, Retzlaff, & Cookson, 1997; see also Kafer & Hunter, 1997). To overcome the latter and, consequently, to enhance its utility as clinical and research tool, Schnirman, Welsh, and Retzlaff (1998) applied an empirical approach to improve the reliability of ToL test scores by selecting from a larger pool of pre-evaluated items those problems that exhibited highest item-total correlations. Thereby, the internal consistency of test scores from the revised ToL problem set could be increased considerably above what would be expected from simply raising the number of items (Cronbach $\alpha = .79$; split-half reliability = .74; Schnirman et al., 1998). However, despite its indisputable merit and the remarkable gain in psychometric quality, the approach of Schnirman et al. (1998) has several constraints. At first, the combination of a relatively rigorous time limit of only 12 s for solution and the uncommon scoring method based on up to three attempts might have placed less emphasis on actual planning ahead, thus possibly favoring less demanding items that might not be optimally suited for detection of interindividual differences in planning ability. Further, an empirical problem selection striving for homogeneity across items may not necessarily result in a linear increase of problem difficulty across different levels of minimum moves, hence questioning again the validity of any interpretations of results derived from the resulting problem set and, in consequence, its suitability for diagnostic purposes. Finally, an empirical approach on problem selection as applied by Schnirman et al. (1998) did not take into account the influence of structural problem parameters other than minimum moves, which, in turn, might lead to a significant loss of relevant diagnostic information. For instance, McKinlay et al. (2008) recently showed that planning impairments in Parkinson's disease can be overlooked easily, as these are related only to a specific aspect of planning. Similar evidence has been obtained in research on the emergence of specific planning abilities in typically developing preschool children (Kaller, Rahm, Spreer, Mader, & Unterrainer, 2008).

Therefore, as an alternative to a purely empirical problem selection, Kaller et al. (2011) suggested a problem set that was derived from comprehensive theoretical analyses of the ToL problem space while taking into account present empirical evidence on the impact of problem structure on planning. The problem set was compiled based on systematic variations of several problem parameters.

The primary aim of the present study was hence to investigate the adequacy of the proposed problem set as a clinical and research instrument. To this end, we assessed (a) whether the present approach yields the predicted linear increase of item difficulty and (b) whether this resulted in a sufficient and coherent separability of individual planning ability that constitutes a basic requirement for diagnostic purposes. In addition, we evaluated the problem set's psychometric properties with specific focus on split-half reliability and internal consistency of the resulting performance scores.

# Method

## Subjects

A total of 130 volunteers (72 male; 12 left-handers; *M* age = 23.7 years, *SD* = 2.9, range = 18.5–32.0) were included in the present analyses. Participants were recruited from undergraduate students of the University of Freiburg. All included subjects had normal or corrected-to-normal vision, and none was under medical treatment or reported a history of neurological or psychiatric disorder. Based on these criteria, another six volunteers were excluded beforehand because of a neurological or psychiatric history and/or intake of psychotropic medication ($n = 5$) or insufficient visual acuity ($n = 1$). Written informed consent was obtained from all subjects prior to the experiment. Subjects received compensation of 10 euros for participation. Present data were acquired as part of a larger study on planning and complex cognition (Kaller & Stahl, 2010).

## Tower of London Task

Subjects were administered a computerized version of the original Tower of London (ToL) consisting of three pegs with different heights. As in most ToL applications, start state and goal state were presented in the lower and upper part of the screen, respectively. Subjects were asked to transform the start state to match the goal state while following three rules: (a) only one ball could be moved at a time; (b) a ball may not be moved if another ball was already on top of it; and (c) three balls could be accommodated at the tallest peg on the left, two balls at the peg in the middle, and one ball at the smallest peg on the right. The computer program did not allow rule-incongruent moves. Movements were executed on a 17-in. (43.18-cm) touch screen.

Individual trials were self-paced and initiated by the subject pressing a button on the screen. Before displaying the next problem, the computer program prompted the subject acoustically to plan ahead first. The minimum number of moves for the current trial was indicated on the left side of the start state. The presentation of each trial was limited to 1 min (cf. Shallice, 1982). Subjects were tested individually. Written as well as verbal instructions placed strict emphasis on always planning ahead the solution before actually moving the balls. Subjects were further asked to complete the task as quickly and accurately as possible,

[1] For convenience, the phrase "minimum moves" hereafter refers interchangeably to the minimum number of moves.

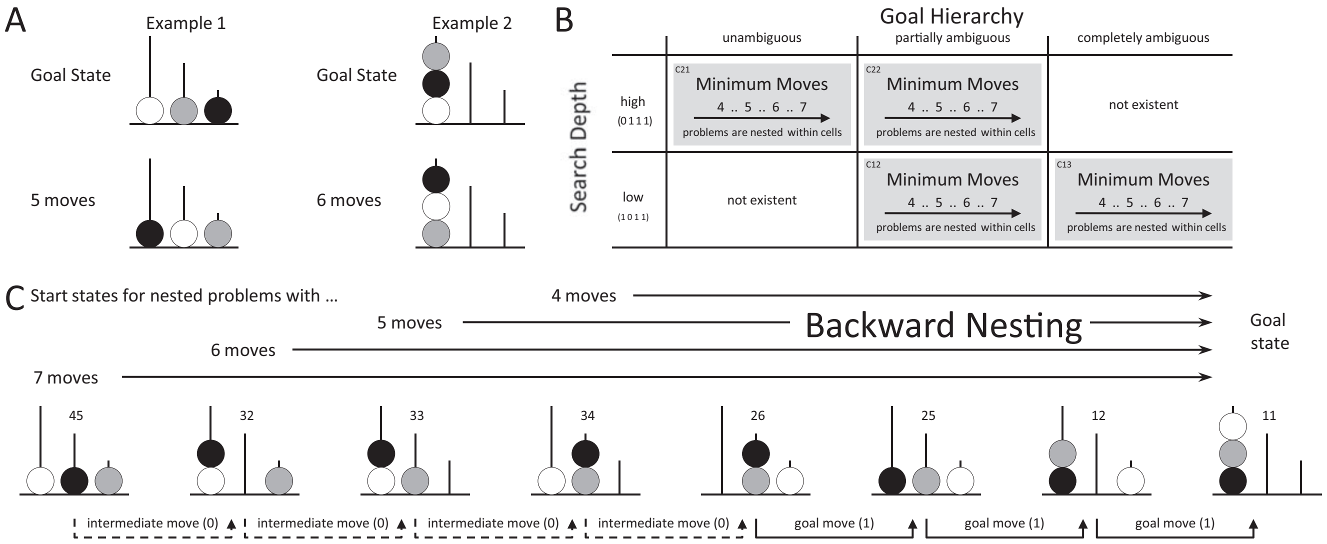

*Figure 1.* A. Examples of two Tower of London problems with minimum numbers of five (left) and six moves (right). Both problems can be accomplished by transforming the respective start state (bottom) into the goal state (top). Although the five-move problem should be easier to solve following the common operational terms of task difficulty, most people find it considerably harder than the illustrated six-move problem. In consequence, considerations of task difficulty solely based on minimum moves fall short of being an adequate approximation. B. Experimental design comprising a factorial manipulation of two structural problem parameters search depth and goal hierarchy across the minimum number of four to seven moves. C. Illustration of the backward nesting strategy across minimum moves with nested problems sharing the last *n* moves. Note that in the actual problem set, nestings were concealed by assigning different permutations of ball colors. Numbers above states refer to the Berg and Byrd (2002) notation.

that is, to solve the problems within the given minimum number of moves.

## Tower of London Problem Set

The applied ToL problem set comprised an extended version of the composition recently suggested by Kaller et al. (2011), ranging from a minimum number of one up to seven moves. Given the sample of presumably high-achieving university students, the inclusion of seven-move problems was intended to provide sufficient discriminability between subjects with an individual planning ability in the upper range. Present analyses are consequently focused on problems with four to seven moves. Note, however, that for other samples (e.g., children, older adults), it may be sufficient to use three- to five-move problems only.

The basic idea behind the construction of the present problem set was to provide a virtually linear increase of difficulty that is tightly linked to the minimum number of moves. As the difficulty of individual ToL problems is substantially determined by other structural problem parameters beside the minimum moves (e.g., Berg et al., 2010; Kaller et al., 2004; Ward & Allport, 1997), it was hence an integral part of the present approach to control for these influences systematically by keeping them at a constant level across the minimum number of moves. Thereby, efforts were focused on the three structurally most eminent ToL problem parameters, that is, goal hierarchy, search depth, and the number of optimal paths to solution (cf. Kaller et al., 2011).

Goal hierarchy is related to the ambiguity of information on subgoal ordering, that is, the degree to which the sequence of final goal moves can be derived from the configuration of the goal state (Kaller et al., 2004; McKinlay et al., 2008; Ward & Allport, 1997). For example, ToL problems with tower goal states, where all three balls are stacked on a single rod, provide an "unambiguous" goal hierarchy since the bottom-most ball has to be placed before the ball that is second from the bottom and so on. In contrast, no such information can be derived from flat goal states where all three balls are distributed on different rods ("completely ambiguous"). Finally, goal states with one and two balls stacked on different rods are "partially ambiguous," as they provide sequential information at least for the two balls lying on top of each other (Kaller et al., 2011).

Search depth constitutes another structural problem parameter that may considerably shift problem difficulty between problems with an identical minimum number of moves (e.g., Kaller et al., 2008). It is defined as the number of intermediate moves before the first ball can be placed into its goal position (Kaller et al., 2011). Intermediate moves are essential to the problem's solution but do not place a ball directly into its goal position (cf. Ward & Allport, 1997). In ToL problems with a minimum solution of four or more moves, search depth refers to mainly two predominant patterns (Kaller et al., 2004): With respect to the last four moves, most problems feature either (a) sequences of one intermediate move followed by three successive goal moves or (b) sequences consisting of a goal move followed by an intermediate move and two successive goal moves (see also Figure 1B). Because all moves before the last four moves are per se intermediate, the two patterns consequently lead to problems with higher and lower search depths.

Finally, the number of optimal paths to solution also has a considerable influence on problem difficulty, because it is easier to find a solution within the minimum number of moves if there is more than one optimal alternative (cf. Berg et al., 2010; Newman & Pittman, 2007; Unterrainer, Rahm, Halsband, & Kaller, 2005). For the present development, the number of optimal paths was therefore maintained constant at only one path to solution, whereas goal hierarchy and search depth were systematically varied across the minimum number of four to seven moves.

Yet because of the general properties of the ToL problem space, a factorial manipulation of goal hierarchy and search depth does not result in a fully balanced design, yielding only four instead of six problem types or cells (see Figure 1B). That is, as certain combinations of the two parameters simply do not exist, comprehensive testing for interactions between them is unfeasible (Winer, 1962). Nonetheless, this factorial approach has two striking advantages. First, as intended, the resulting problem set will presumably permit to attain the linear increase of problem difficulty across minimum moves. Second, as a side effect, the problem set may even provide more specific information on the causes of altered planning ability as was recently demonstrated for planning impairments in Parkinson's disease patients (McKinlay et al., 2008; for other examples, see Kaller et al., 2011).

For practical reasons, it is, however, nearly impossible to systematically control for all structural problem parameters that may have an impact on planning (for an overview, see Kaller et al., 2011). To further restrain potential influences of other factors, the construction of the present problem set was based on a backward-nesting strategy. That is, paths of problems with less minimum moves were nested into the solution paths of problems with higher numbers of minimum moves. More specifically, two nested problem families were selected for each of the four problem types or cells of the factorial design (see Figure 1B). Each of these problem families consisted of a four-, a five-, a six- and a seven-move problem that were backward-nested into each other based on the sequences of the respective last four to seven moves (see Table 1). Concerning other structural aspects, problems within cells were hence equalized as close as possible across the minimum number of moves. Backward-nesting denotes that all problems structurally shared the last n moves, although this was concealed by different permutations of ball colors. For sake of clarity, the basic principle of this nesting strategy is illustrated in Figure 1C, using problem paths with identical permutations of ball colors (cf. Berg & Byrd, 2002).

For the final set, problems were selected in a pseudorandom manner from a pool of structurally identical iso-problems with different permutations of ball colors to control for and minimize the overlap of single move sequences within as well as between nested problem families. To further avoid carryover effects due to repeated presentations of nested problems, the presentation of problems was in a fixed order of problem types within successively increasing levels of minimum moves (see Table S1 in the online supplemental materials). Within a given minimum move level, the assignment of the two nested problem families to the first or second incidence of the respective problem type was also pseudorandomly controlled to minimize the overlap of move sequences between neighboring trials, irrespective of ball colors. Nevertheless, this resulted in a relatively systematic order of nested problem families across minimum moves.

The final problem set as applied in the present study together with comprehensive information concerning its structural properties are listed in supplemental Table S1. Interested readers are also referred to Kaller et al. (2011) and the accompanied open-source software TowerTool for in-depth visualizations of the suggested problem set and further details on problem structure (http://www.uniklinik-freiburg.de/fbi/live/apps/towertasks.html).

## Measures

Accuracy of problem solving as well as several latency measures, such as initial thinking and movement execution times, were recorded. For the present analyses on planning ability and psychometric properties of the problem set, however, only the accuracy data were considered. Accuracy indicated whether a problem was correctly solved in the minimum number of moves or not.

# Results

## Linear Increase of Problem Difficulty

Meaningful assessment of planning ability in clinical and research settings demands a problem composition that allows to test and discriminate subjects across a broad range of individual capa-

Table 1
*Bases of the Problem Set*

| | Parameter | | Start State:Goal State in Berg & Byrd (2002) notation | | | |
|---|---|---|---|---|---|---|
| Base | *F*(SD) | *F*(GH) | 4 Moves | 5 Moves | 6 Moves | 7 Moves |
| A | high | low | 14:51 (34:11) | 23:41 (33:11) | 12:51 (32:11) | 55:21 (45:11) |
| B | high | low | 45:51 (65:11) | 52:11 (52:11) | 53:11 (53:11) | 24:61 (54:11) |
| C | high | high | 23:64 (53:14) | 34:54 (54:14) | 56:24 (46:14) | 25:54 (45:14) |
| D | high | high | 14:53 (34:13) | 43:63 (33:13) | 42:63 (32:13) | 15:43 (45:13) |
| E | low | low | 16:24 (26:14) | 54:34 (34:14) | 53:34 (33:14) | 32:14 (32:14) |
| F | low | low | 22:63 (52:13) | 63:43 (53:13) | 24:63 (54:13) | 46:13 (46:13) |
| G | low | high | 46:65 (36:15) | 15:55 (35:15) | 42:15 (42:15) | 63:35 (43:15) |
| H | low | high | 42:55 (62:15) | 45:25 (55:15) | 16:35 (56:15) | 24:55 (44:15) |

*Note*. SD = search depth; GH = goal hierarchy; *F*(SD) = transformation into factorial design (low vs. high SD); *F*(GH) = transformation into factorial design (low vs. high ambiguity). To illustrate the backward nesting within bases, problem states in parentheses share identical color permutation for the goal state.

bilities. Thus, a basic requirement for the present ToL problem set was a continuous increase of task difficulty across the minimum number of moves. To evaluate this property, a one-way repeated-measurements analysis of variance (RM-ANOVA) of the accuracy data was computed across the four levels of minimum moves (i.e., four to seven moves). Results revealed a significant main effect, $F(3, 387) = 303.96$, $p < .001$, that was further substantiated by significant contrasts for comparisons between subsequent difficulty levels: four vs. five moves, $F(1, 129) = 126.48$, $p < .001$; five vs. six moves, $F(1, 129) = 82.79$, $p < .001$; six vs. seven moves, $F(1, 129) = 65.94$, $p < .001$. As displayed in Figure 2A, mean accuracy decreased gradually by nearly 20% for each increase in the minimum number of moves.

In addition, despite the presumably rather homogeneous and high-achieving sample of university students, subjects' individual performance ranged from 18.75% to 87.50% ($M = 57.91$, $SD = 13.15$) thereby following a Gaussian distribution (see Figure 2B; Kolmogorov-Smirnov test, $p = .334$). As is evident from Figure 2C, the linear increase of problem difficulty across the minimum number of moves was not just a simple average group phenomenon but was found also to correspond highly with the individual limits of subjects' planning ability. That is, low-, medium- and high-performing subjects were likely to attain correct solutions in problems up to a level of low-, medium-, and high-minimum moves, respectively, but not above. This pattern was also to some extent reflected in the correlation structure of the minimum moves subscales amongst each other as well as with the part–whole corrected or total overall score (see Table 2). As performance in four- and seven-move problems was approaching ceiling and floor, respectively, the subscales with the medium difficult five- and six-move problems had, at least for the given sample, higher discrimination characteristics (as indicated by the part–whole correlations) and, consequently, showed stronger contributions to the between-subjects variability in estimated planning ability (as indicated by the correlations with the total overall score). On the other hand, the general magnitude of correlations was only low to moderate, presumably because of the broad range of item difficulty in the problem set (see Table 2).

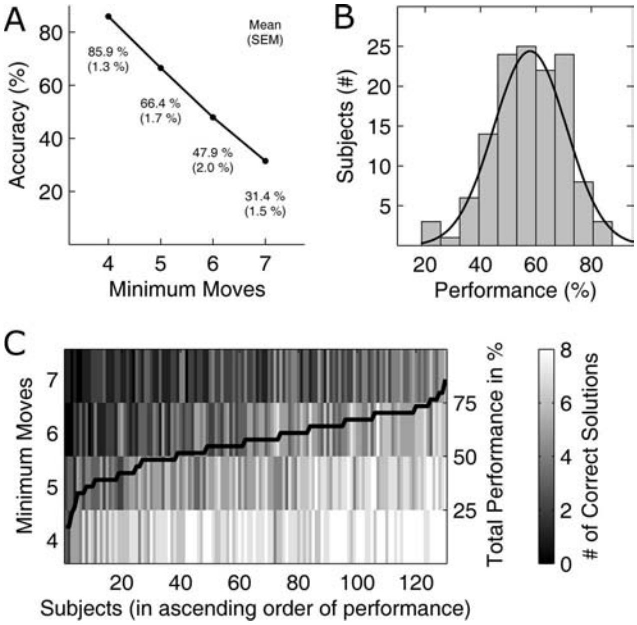

*Figure 2.* A. Accuracy as a function of the minimum number of moves yielding a linear increase of problem difficulty. B. Performance distribution across subjects approximating normality. C. Performance of individual subjects with respect to problem difficulty in terms of minimum moves. Solid line indicates subjects' total performance, whereas black-to-white patchings denote individual number of correct solutions per minimum moves.

Table 2
*Correlation Analyses of Minimum-Moves Subscales*

| | Correlation with subscale scores | | | | Overall correlation | |
|---|---|---|---|---|---|---|
| Subscale | 4 Move | 5 Move | 6 Move | 7 Move | P/W | Total |
| 4 Move | 1.00 | .373 | .301 | .284 | .425 | .641 |
| 5 Move | | 1.00 | .392 | .297 | .482 | .742 |
| 6 Move | | | 1.00 | .341 | .471 | .767 |
| 7 Move | | | | 1.00 | .409 | .663 |

*Note.* P/W = part-whole correlation. All *p*s < .01.

## Split-Half Reliability, Internal Consistency and Precision of Alpha

Split-half reliability of test scores was estimated using a twin-items approach based on pairs of the two problems with equal structural properties at each level of minimum moves. Instead of a single-shot random allocation of each item into one half, an exhaustive permutation approach was used by computing correlation indices for all $2^{16}$ possible assignments of item pairs. Resulting correlations were adapted using the Spearman-Brown prophecy formula (cf. Cortina, 1993). Results revealed an average split-half reliability of $r = .715$ and an estimated maximum reliability of $r = .826$. Applying sample-size corrections suggested by Kristof (1963) resulted in a virtually identical estimation of the average ($r = .718$) and maximum split-half reliability ($r = .828$).

Because item responses were scaled dichotomically (correctly solved in the minimum number of moves: true/false), estimation of internal consistency was based on the Kuder-Richardson precursor of the Cronbach (1951) formula. Interrelatedness of items as represented by the alpha coefficient was at $\alpha = .691$ and thereby close to the average split-half reliability, as would be expected given that twin-paired items showed comparable standard deviations (see supplemental Table S1; cf. Cortina, 1993). Because of the intended variations in item difficulty across the minimum four to seven moves and the resulting differences in item discrimination indices (cf. part–whole point-biserial correlations in supplemental Table S1), the alpha coefficient hence represents the lower bound of reliability, rather than an exact estimate. Furthermore, because the alpha coefficient does not per se imply unidimensionality or homogeneity (Cortina, 1993), the precision of alpha in terms of the standard error of interitem correlations was also considered. Although mean interitem correlations were found to be low ($r = .064$), the precision of alpha was also close to zero ($P_\alpha < .001$) and thus did not indicate an instance of critical heterogeneity within the problem set.

### Item Bias Analyses

To test for item biases, the sample was odd-even split into two halves for separate calculations of item difficulty. Ordinal item ranks were then determined and correlated between halves using Spearman's rank correlation. Results revealed a correlation of $r = .944$, indicating a high correspondence of item difficulties estimated separately for the two random subsamples.

Robustness of item difficulties was further corroborated by additional analyses based on data of another sample of subjects independently assessed by Berg et al. (2010). In this comprehensive study, Berg and colleagues empirically explored the item characteristics of virtually all ToL problems with moderate to high levels of difficulty. Most notably, Spearman's rank correlation revealed also a high correspondence of $r = .931$ between the item difficulties estimated in the present study and the item difficulties of the structurally identical problems[2] acquired by Berg et al. (2010).

### Supplementary Analyses

All preceding analyses were focused on the psychometric properties of the problem set as a whole and its subscales (in terms of minimum moves). Readers interested in further in-depth analyses on the characteristics of individual items are referred to the online supplemental materials that also provide single-item based measures of dispersion, time-out records and discrimination indices in terms of the part–whole point-biserial correlations with the overall and subscale scores. Analyses on the impact of problem structure across the minimum of four to seven moves are also reported in the supplemental materials.

## Discussion

In the present study, we investigated the psychometric properties of a set of ToL problems that was composed on the basis of theoretical problem space analyses (Kaller et al., 2011). Development and evaluation of this problem set were based on several intentions. First and foremost, by a factorial manipulation of different structural problem parameters that are known to influence the assessment of planning ability (e.g., Berg et al., 2010; Kaller et al., 2004, 2008; McKinlay et al., 2008; Newman & Pittman, 2007; Unterrainer et al., 2005; Ward & Allport, 1997; for an overview, see Kaller et al., 2011), we strived to develop a consistent problem set that (a) is well suited for diagnostic purposes by effectively realizing a linear increase of planning demands along the minimum number of moves and (b) may yield more in-depth information with respect to specific aspects of planning (e.g., Kaller et al., 2008; McKinlay et al., 2008). Second, by assessing its psychometric quality, we intended to evaluate the problem set's general suitability for an application in clinical and research contexts. Third and finally, by providing descriptive and statistical values for every single item (see supplemental Table S1), we aimed to facilitate further improvements of the suggested problem selection toward a widely accepted ToL standard problem set that finally ensures the often requested but still lacking comparability of results across studies and research groups (cf. Berg & Byrd, 2002; Hinz et al., 2009; Kaller et al., 2004, 2011; Sullivan et al., 2009).

Consistent with the study's intentions, results showed a clear and nearly perfect linear increase of task difficulty across minimum moves (see Figure 2A). Specifically, mean accuracy decreased gradually by 20% per difficulty level. Persuasive validity of the test scores and, in turn, appropriateness of the operationalization of the applied problem set was also demonstrated on the individual level as subjects' overall performance was closely reflected by their respective achievements across different levels of difficulty (see Figure 2C). In other words, low-performing subjects failed gradually to solve problems at more demanding levels, whereas high-performing subjects reliably solved easier problems. In addition, despite the presumably rather homogeneous sample of university students, the problem set allowed us to discriminate well between individual planning abilities. However, ceiling and floor effects became apparent at least to some extent in four- and seven-move problems, respectively. But given its intended use primarily for diagnostic purposes, the wide range of task difficulty enables researchers and clinicians to apply a single problem set to assess planning ability across a variety of potential subjects ranging from preschool children and neuropsychological patients to university students and putative planning experts, such as chess players.

Furthermore, split-half reliability and internal consistency of the suggested problem set were acceptable (.715 and .691, respectively) and far superior to the original ToL set (.19 and .25; cf. Humes et al., 1997), even after adjusting for the smaller number of the original 15 items compared with the present 32 items using the Spearman-Brown prophecy formula (.33 and .42, respectively). Although Schnirman et al. (1998) attained a comparable split-half reliability (.74) but slightly higher internal consistency (.79), this latter discrepancy can be readily explained by the two-stage item selection process. That is, Schnirman and colleagues assessed 69 problems in a first stage and subsequently calculated item-total correlations to identify the 30 items with the highest correlations with the overall score. The reduced problem set was then assessed in a second stage, resulting in an inherently higher estimate of the test scores' internal consistency. However, as noted before, this purely empirical approach does not account for structural properties of the ToL that were shown to have a substantial impact on whether planning disturbances can be reliably detected (cf. Kaller et al., 2008; McKinlay et al., 2008, 2009). Furthermore, an empirical item selection is always constrained by the underlying item pool and the drawn sample of subjects, which is not the case for a theoretical approach based on structural task analyses (Kaller et al., 2004, 2011). Thus, we believe that the evident advantages of the present approach clearly outweigh a slightly lower internal consistency, particularly because the present comprehensive and transparent documentation of single-item characteristics allows for targeted improvements in future revisions.

In supplementary analyses, previously established effects of problem structure were replicated (e.g., Berg et al., 2010; McKinlay et al., 2008; see the online supplemental materials). These

[2] We are indebted to Keith Berg for kindly providing us with this data. For detailed information on sample characteristics and experimental procedures, please refer to Berg et al. (2010).

findings suggest that the present set of ToL problems can be used not only to measure overall planning ability but also to reveal more detailed information about planning ability on the level of individual subjects or specific samples (for review, see Kaller et al., 2011). Recent research suggests that planning ability is a complex construct and as of yet not fully understood (Koppenol-Gonzalez, Bouwmeester, & Boonstra, 2010). Although it is beyond doubt that structural problem parameters such as minimum moves, goal hierarchy, and search depth are important determinants of performance (e.g., Kaller et al., 2004, 2011; Newman & Pittman, 2007; Ward & Allport, 1997), other properties of individual problems appear also to have an influence on planning accuracy (cf. Berg et al., 2010). Future studies should therefore address these task parameters and highlight their specific impact on planning in the general context of other contributing factors. In this respect, a variety of possible research strategies was recently discussed by Kaller et al. (2011). In addition, several suggestions for further improvements of the present problem set toward the development of a final ToL standard set are provided in the online supplemental materials.

Meanwhile, the present study showed that measurement of planning ability in clinical and research settings can already be improved substantially. Taken together, the ToL problem set evaluated here proved to have good psychometric qualities and to provide a conceptually sound basis for the diagnosis of interindividual differences by unfolding a nearly linear increase of task difficulty across the minimum number of moves. Because of the factorial manipulations of problem structure, the problem set further opens promising new perspectives on more specific diagnostics of planning disturbances. In a recent meta-analysis on various tower tasks, Sullivan et al. (2009) outlined that tasks like the Towers of London, Hanoi and their variants are able to detect impaired performance in various neuropsychological patient samples. However, in close parallel with the spirit of the present article, Sullivan et al. critically concluded that it is not sufficient to explore only the tasks psychometric properties but that future research has to demonstrate which variables of tower tasks are measuring which particular aspect of planning, as well as delineate a specific involvement of circumscribed neural substrates. We hope that the present analyses form a first step toward this goal.

## References

Berg, W. K., & Byrd, D. (2002). The Tower of London spatial problem-solving task: Enhancing clinical and research implementation. *Journal of Clinical and Experimental Neuropsychology, 24,* 586–604. doi:10.1076/jcen.24.5.586.1006

Berg, W. K., Byrd, D. L., McNamara, J. P. H., & Case, K. (2010). Deconstructing the tower: Parameters and predictors of problem difficulty on the Tower of London task. *Brain and Cognition, 72,* 472–482. doi:10.1016/j.bandc.2010.01.002

Carder, H., Handley, S., & Perfect, T. (2004). Deconstructing the Tower of London: Alternative moves and conflict resolution as predictors of task performance. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 57A,* 1459–1483. doi:10.1080/02724980343000864

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98–104. doi:10.1037/0021-9010.78.1.98

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334. doi:10.1007/BF02310555

Goel, V. (2002). Planning: Neural and psychological. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (Vol. 3, pp. 697–703). London, England: Nature Publishing Group.

Hinz, A. M., Kostov, A., Kneißl, F., Sürer, F., & Danek, A. (2009). A mathematical model and a computer tool for the Tower of Hanoi and Tower of London puzzles. *Information Sciences, 179,* 2934–2947. doi:10.1016/j.ins.2009.04.010

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

Kafer, K. L., & Hunter, M. (1997). On testing the face validity of planning/problem-solving tasks in a normal population. *Journal of the International Neuropsychological Society, 3,* 108–119.

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. doi:10.1016/j.bbr.2010.07.029

Kaller, C. P., Rahm, B., Spreer, J., Mader, I., & Unterrainer, J. M. (2008). Thinking around the corner: The development of planning abilities. *Brain and Cognition, 67,* 360–370. doi:10.1016/j.bandc.2008.02.003

Kaller, C. P., & Stahl, C. (2010). *Investigating the cognitive foundations of planning abilities.* Unpublished data.

Kaller, C. P., Unterrainer, J. M., Rahm, B., & Halsband, U. (2004). The impact of problem structure on planning: Insights from the Tower of London task. *Cognitive Brain Research, 20,* 462–472. doi:10.1016/j.cogbrainres.2004.04.002

Koppenol-Gonzalez, G. V., Bouwmeester, S., & Boonstra, A. M. (2010). Understanding planning ability measured by the Tower of London: An evaluation of its internal structure by latent variable modeling. *Psychological Assessment, 22,* 923–934. doi:10.1037/a0020826

Kristof, W. (1963). Die Verteilung aufgewerteter Zuverlässigkeitskoeffizienten auf der Grundlage von Testhälften [The distribution of reliability coefficients based on split-half approaches]. *Archiv für die gesamte Psychologie, 115,* 230–240.

McKinlay, A., Grace, R. C., Kaller, C. P., Dalrymple-Alford, J. C., Anderson, T. J., Fink, J., & Roger, D. (2009). Assessing cognitive impairment in Parkinson's disease: A comparison of two tower tasks. *Applied Neuropsychology, 16,* 177–185. doi:10.1080/09084280903098661

McKinlay, A., Kaller, C. P., Grace, R. C., Dalrymple-Alford, J. C., Anderson, T. J., Fink, J., & Roger, D. (2008). Planning in Parkinson's disease: A matter of problem structure? *Neuropsychologia, 46,* 384–389. doi:10.1016/j.neuropsychologia.2007.08.018

Newman, S. D., & Pittman, G. (2007). The Tower of London: A study of the effect of problem structure on planning. *Journal of Clinical and Experimental Neuropsychology, 29,* 332–342. doi:10.1080/13803390701249051

Owen, A. M. (2005). Cognitive planning in humans: New insights from the Tower of London (TOL) task. In R. Morris & G. Ward (Eds.), *The cognitive psychology of planning* (pp. 135–151). Hove, England: Psychology Press.

Schnirman, G. M., Welsh, M. C., & Retzlaff, P. D. (1998). Development of the Tower of London—Revised. *Assessment, 5,* 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. doi:10.1098/rstb.1982.0082

Sullivan, J. R., Riccio, C. A., & Castillo, C. L. (2009). Concurrent validity of the tower tasks as measures of executive function in adults: A meta-analysis. *Applied Neuropsychology, 16,* 62–75. doi:10.1080/09084280802644243

Unterrainer, J. M., Rahm, B., Halsband, U., & Kaller, C. P. (2005). What is in a name: Comparing the Tower of London with one of its variants. *Cognitive Brain Research, 23,* 418–428. doi:10.1016/j.cogbrainres.2004.11.013

Ward, G., & Allport, A. (1997). Planning and problem-solving using the five-disc Tower of London task. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 50,* 49–78. doi:10.1080/027249897392224

Ward, G., & Morris, R. (2005). Introduction to the psychology of planning. In R. Morris & G. Ward (Eds.), *The cognitive psychology of planning* (pp. 1–34). Hove, England: Psychology Press.

Winer, B. J. (1962). *Statistical principles in experimental design.* New York, NY: MacGraw-Hill. doi:10.1037/11774-000

Received March 16, 2011
Revision received July 5, 2011
Accepted July 6, 2011 ■