VIENNA TEST SYSTEM

MANUAL

TOWER OF LONDON - FREIBURG VERSION

Test label TOL-F

Version 21

Mödling, October 2011

Authors of the test Christoph P. Kaller, Josef M. Unterrainer, Stefan Kaiser, Matthias Weisbrod & Steffen Aschenbrenner
Authors of the manual Christoph P. Kaller, Josef M. Unterrainer, Stefan Kaiser, Matthias Weisbrod, Rudolf Debelak & Steffen Aschenbrenner

SCHUHFRIED GmbH, Hyrtlstraße 45, 2340 Mödling, Austria
Tel. +43/2236/42315-0, Fax: +43/2236/46597
info@schuhfried.at www.schuhfried.at
Sitz: Mödling, FN 104661p
Landesgericht Wr. Neustadt, UID Nr. ATU 19273809

SCHUHFRIED
passion for psychology

CONTENTS

1 SUMMARY

Application

Test to measure planning ability in healthy individuals and in psychiatric and neurological patients.

Theoretical background

The term “planning ability” is used here to describe the ability to model solution possibilities cognitively and to assess the consequences of an action before it is carried out. The “Tower of London” dates back to an attempt by Shallice (1982) to devise a planning task that covers a broad difficulty spectrum and hence makes it possible to administer a large number of qualitatively different problems. The present version is based on the findings of recent studies of the connection between task complexity and the cognitive processes that underlie planning ability. Use of the TOL-F is recommended for various neurological disorders (e.g. frontal brain injury, neurodegenerative diseases) and psychiatric disorders (e.g. schizophrenia, compulsive disorders) in which planning ability is likely to be impaired.

Administration

The present test provides a detailed evaluation of planning ability and hence enables a precise assessment, which can be used as a basis for therapeutic intervention. Either the standard or the short form of the TOL-F can be used, depending on the reason for the investigation and the patient’s ability level.

Test forms

There are two test forms. The first form is the standard form, which provides a detailed assessment of planning ability. The second form is a short form which discriminates mainly in the lower ability range; it therefore enables quick and economical measurement of performance deficits. Both the standard and the short forms of the TOL-F are available in three parallel versions.

Scoring

The main target variable is *planning ability* – i.e. the number of items worked correctly within a time limit of one minute each. Information on error types (such as systematic rule infringements or changes of mind while working the items) and on planning and execution times is reported.

Reliability

The test’s reliability was estimated from the data of the norm sample. Cronbach’s Alpha and other measures of reliability for planning ability as the main variable are >0.7 and thus – bearing in mind the broad range of different item difficulties combined with the relatively short test duration – are entirely satisfactory.

Validity

Extensive literature supports the validity of the test implemented here. Variants of the “Tower of London” had already been used with numerous neurological and psychiatric patient groups and with healthy adults and children. The present variant is based on a number of recent studies of the psychometric properties of the “Tower of London”.

Norms

Data is available for 269 individuals from the normal population aged between 16 and 84 years, distributed approximately uniformly with regard to age and gender.

Time required for the test

The standard form of the TOL-F takes around 16 minutes to complete; the short form requires around 11 minutes.

2 DESCRIPTION OF THE TEST[1]

2.1 Theoretical background

2.1.1 Planning and problem solving

Human behavior is characterized in many situations by a focus on a particular goal. Outside everyday routines, however, intended goals can often not be achieved directly because steps that are necessary for attainment of the goal are dependent on each other or even mutually exclusive and familiar action schemata cannot be applied. Such situations, in which existing and often overlearned reaction patterns cannot be applied or do not lead to a satisfactory outcome, present problems that can be solved in various ways. Problem-solving behavior in the strict sense has three distinct properties: (1) a focus on a goal, (2) the need to break down the desired goal into sub-goals and (3) the use of operators, i.e. actions that convert the problem from an existing state into a new one (Anderson 2005). It follows that solving a problem often involves finding an appropriate sequence of actions for the given situation. In the simplest case this can be done by trial and error; for efficient behavior, however, it is often necessary to use prior planning to identify an action sequence that is conducive to attainment of the goal.

Planning is thus a form of problem solving; it involves the mental simulation and evaluation of action sequences and the resulting consequences (Goel 2002). The ability to plan makes it possible to organize goal-directed behavior before it is actually carried out in the dimensions of time and space (Owen 1997) and to select from a range of behavior options on the basis of the modeled prospects of success (Ward & Morris 2005).

Problem situations therefore arise from lack of awareness of a transformation function by means of which a given starting state can be converted into a desired goal state through the use of the available operators while taking account of the existing restrictions. In addition, problems can be classified in terms of the available information (Goel 2002). Challenges in everyday life are usually under-specified or ill-defined, i.e. the starting and/or finishing state is described unclearly or not at all, as are the possible operators and restrictions (Ormerod 2005). For example, the plan to cook a meal for guests does not specify how hungry the guests will be or how much effort is to be put into put into preparing the meal. Moreover, the goal state is only vaguely formulated; it is not stipulated whether one should serve three courses or four, what choice of foods would be appropriate and to what extent one is trying to impress one’s guests. The possible operators are also unclear: the options include cooking the meal oneself, ordering from a catering service or asking each of the guests to bring a component of the meal (see Goel & Grafman 2000). Poorly defined problems are distinguished from closed, well-defined ones, in which the start and goal states, and the operators and restrictions, are clearly identifiable (Davies 2005). For example, if operating a light switch does not make the light come on because one of several light bulbs is broken, the problem is usually a clearly defined one (see Knoblich & Öllinger 2008). A further distinction is made between problems according to whether finding a solution requires prior knowledge that goes beyond the given situation (Ward & Morris 2005).

In clinics and research, planning ability is usually investigated using well-defined problems that contain all the information necessary for their solution, require no prior knowledge and can be solved unambiguously by identifying a sequence of action steps (for a critical discussion see e.g. Burgess, Simons, Coates & Channon 2005). Planning processes in this context are often described as a look-ahead search for an optimum pathway within an abstract state or problem space that consists of the possible states and the associated operators (Newell & Simon 1972). However, because of the capacity limits of human information processing, this search is seldom comprehensive but instead is limited by

[1] Note: This description is based in part on Kaller (2010).

strategies and heuristics. Often, too, solutions are not completely but only partially pre-planned, so that a distinction is made between initial planning and planning during the execution phase (concurrent planning); the two forms of planning are not mutually exclusive (Davies 2005). Planning behavior is determined in the main by three factors: the complexity or difficulty of the problem, (2) external properties of the problem-solving environment and (3) intra- and inter-individual differences (Davies 2005). For example, there is evidence that initial planning leads to success only for problems of easy to moderate difficulty (Davies 2003). Initial planning is directly favored by explicit instructions (Unterrainer, Rahm, Leonhart, Ruff & Halsband 2003) and also by the extent to which the rules to be observed are externally represented in the problem-solving environment, so that remembering them makes fewer demands on limited processing resources (Kotovsky, Hayes & Simon 1985; Zhang & Norman 1994). Similarly, planning behavior varies intra-individually with the level of experience (Anzai & Simon 1979) and inter-individually between different clinical and non-clinical groups (see Davies 2005).

2.1.2 Disk transfer tasks and the “Tower of London”

A number of different paradigms for measuring planning ability have been developed. Prominent among tests of this type are so-called disk transfer tasks, of which the “Tower of Hanoi” and the “Tower of London” are the best-known examples (see Figure 2.1).The “Tower of Hanoi” was originally invented by Édouard Lucas as a mathematical game and published under an anagram of his name (Claus, 1884, cited in Hinz, Kostov, Kneißl, Sürer & Danek 2009). The task became established in cognitive psychology research largely through the work of Allen Newell and Herbert A. Simon (1972) On the basis of the “Tower of Hanoi”, Tim Shallice (1982) developed the “Tower of London” as a more suitable planning test for patients with frontal lesions.

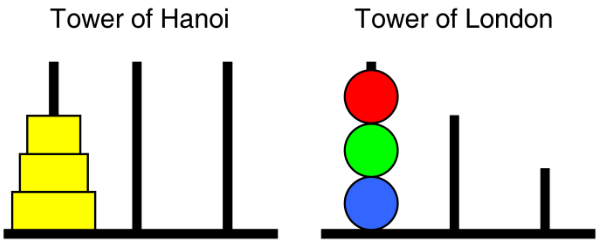


Figure 2.1 Schematic representation of the “Tower of Hanoi” and “Tower of London” (from Kaller, Rahm, Köstering & Unterrainer 2011a).

The aim of both tasks is to convert a starting state into a defined finishing state in as few moves as possible. Various rules must be observed in the process, so that an optimum solution can usually be achieved only with prior planning (but on this point see Simon 1975). For example, only one of the objects (disks or balls) may be moved at a time, and an object may not be placed anywhere other than on the rods. It should also be noted that in the “Tower of Hanoi” a larger disk must never be placed on a smaller one; in the “Tower of London”, by contrast, the balls are of equal size but restrictions are imposed by the height and hence the capacity of the rods (one, two or three balls; see Figure 2.1). Both tests are often used in investigating planning ability (Berg & Byrd 2002). In the clinical and cognitive neurosciences, however, the “Tower of London” has become more widespread than the “Tower of Hanoi”, with more than 200 publications referring to it listed in MEDLINE by the end of 2009 (Figure 2.2).

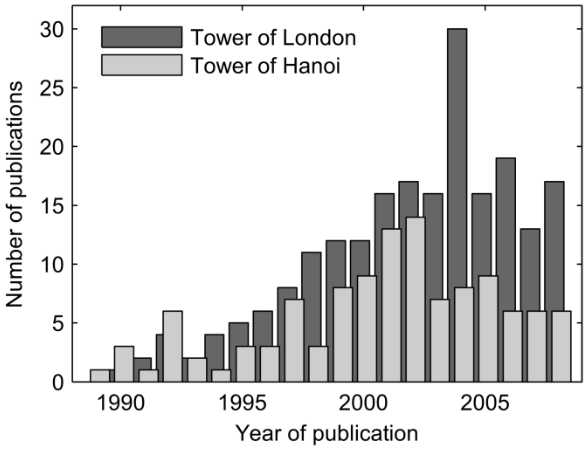


Figure 2.2 Number of publications including the keywords “Tower of London” and “Tower of Hanoi” listed in MEDLINE (http://www.ncbi.nlm.nih.gov/pubmed) between 1990 and 2009. In this 20-year period there are 236 references in MEDLINE to the “Tower of London” and 122 to the “Tower of Hanoi”, with 8 publications referring to both (from: Kaller et al. 2011a).

2.1.3 Approaches to operationalizing planning difficulty

Despite the popularity and widespread use of the “Tower of London” to measure planning ability, very few studies have explored the underlying cognitive processes in detail (see Kaller, Unterrainer, Rahm & Halsband 2004). Instead, the term “Tower of London” is used to cover a wide range of modifications and variants of the task, which may make very different cognitive demands on the respondent (for an overview see Berg & Byrd 2002). The problems used also tend to be selected on an undifferentiated basis, so that it is questionable whether all the studies measure a homogeneous and comparable construct (Kaller et al. 2004; Sullivan et al. 2009). Further support for this view comes from psychometric studies that have found low split-half reliabilities and internal consistencies in the tests used (Humes, Welsh, Retzlaff & Cookson 1997; Schnirman, Welsh & Retzlaff 1998) as well as unsatisfactory construct validity (Kafer & Hunter 1997).

In the majority of studies with the “Tower of London” it is implicitly or explicitly assumed that problem difficulty is determined mainly by the minimum number of moves. However, this assumption is over-simplified: as the comparison of the two five-move problems shown in Figure 2.3 A and B shows, the cognitive processes that underlie planning performance are also determined to a significant extent by other structural properties of the problems to be solved (see Berg, Byrd, McNamara & Case 2010; Kaller et al., 2004; Ward & Allport 1997). As another argument against the widespread operationalization of task difficulty solely in the form of the minimum number of moves, it can be shown that problems that require a higher number of moves are not necessarily more difficult to solve than problems with a lower number of moves (see Figure 2.3 C and also McKinlay et al. 2008).

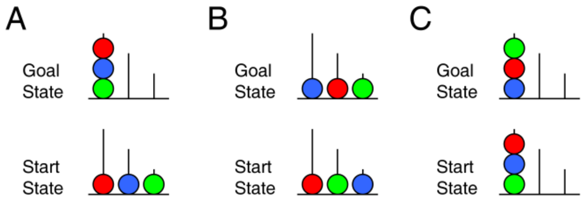
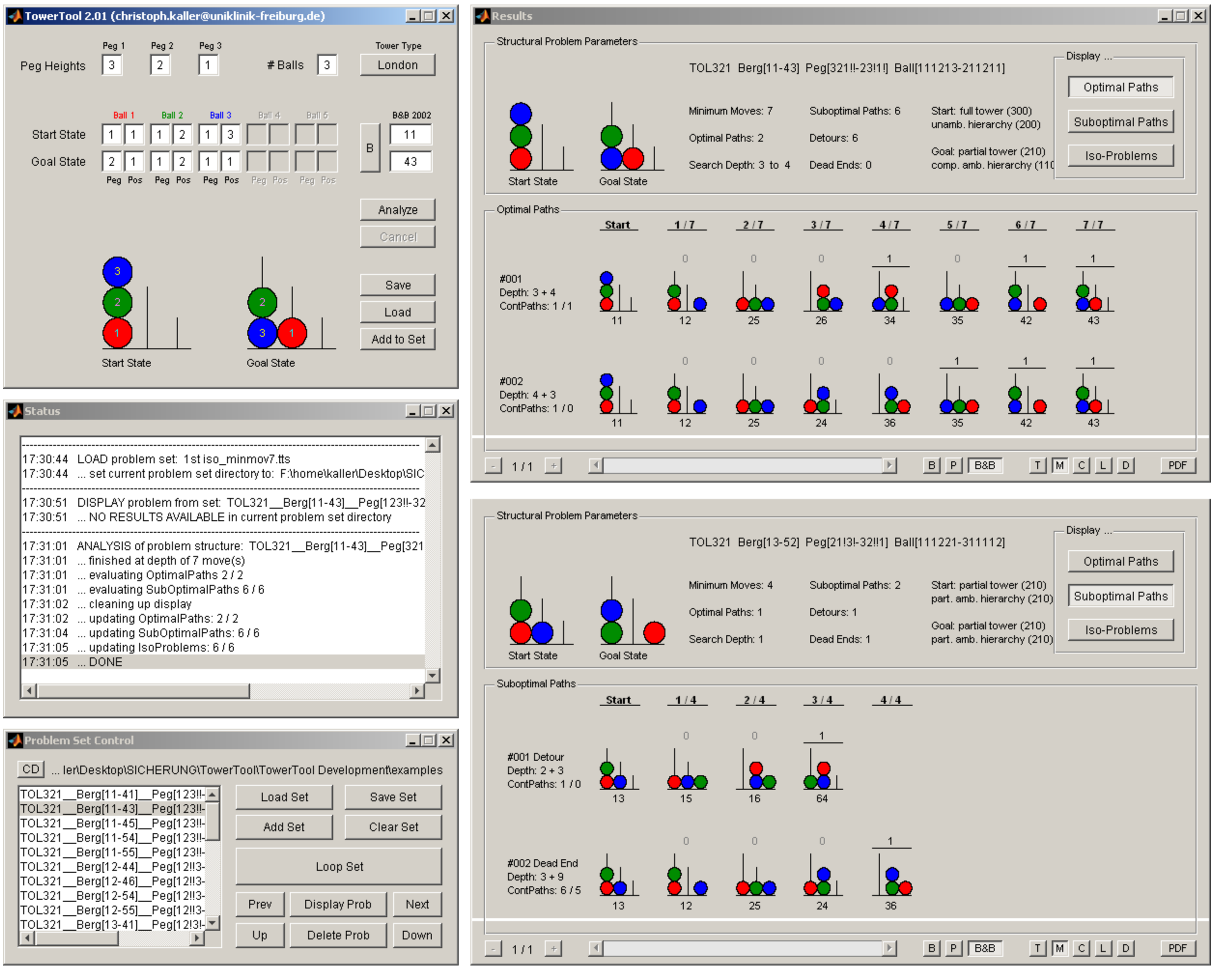


Figure 2.3 Examples of “Tower of London” problems with five (A, B) and six (C) moves. For many respondents the problem shown in (A) is significantly easier than that shown in (B), despite the fact that the minimum number of moves needed to arrive at a solution is the same in both cases (five). The six-move problem shown in (C) is likewise often found to be easier than (B), even though in accordance with the usual operationalization of task difficulty it should be more difficult (based on Kaller et al. 2011a, 2012a, where further details can also be found).

A complex phenomenon such as planning ability cannot be satisfactorily measured with imprecise and over-simplified operationalizations. In addition, even studies carried out some time ago with the “Tower of Hanoi” demonstrate the influence of individual structural problem parameters on planning performance (Borys, Spitz & Dorans 1982; Klahr & Robinson 1981; Spitz, Webster & Borys 1982). On the basis of theoretical considerations and detailed task analysis involving the problem space of the original version of the “Tower of London”, Kaller et al. (2004) therefore isolated various structural problem parameters and tested their effect on initial planning time as an indicator of underlying cognitive processes. By controlling simultaneously for other influences they succeeded in showing for the first time that various task parameters exercise a systematic influence on initial planning time and hence on the cognitive processes involved in the planning processes; they do this independently of each other and even within very simple problems involving only three moves (Kaller et al. 2004; see also Berg et al. 2010). In addition, Unterrainer, Rahm, Halsband and Kaller (2005) showed that in more complex problems, too, planning processes are determined to a significant extent by structural problem parameters; this is true even without regard to global differences between the problem spaces of non-isomorphic tower variants. When task-specific structural parameters are identical, the original “Tower of London” and non-isomorphic variants have the same problem difficulty. By contrast, differences in problem parameters lead to differences in the cognitive demands of different tower variants, even if the external appearance of the individual problems is similar (Unterrainer et al. 2005).

It can therefore be assumed that, when exploring and describing differences in the planning ability of clinical and non-clinical samples and decoding the cognitive and neuronal basis of planning ability, taking structural problem parameters into account is likely to be highly informative. For example, McKinlay et al. (2008) showed using the “Tower of London” that Parkinson’s patients do not suffer from general impairment of their planning ability but that planning deficits are associated with specific task requirements (see also Köstering, McKinlay, Stahl & Kaller, submitted). In the area of cognitive development, too, age-related changes in certain planning-related tasks can be identified and isolated through systematic manipulation of structural problem parameters (Kaller, Rahm, Spreer, Mader & Unterrainer 2008; Klahr & Robinson 1981; Spitz, Webster & Borys 1982). This applies not only to development in childhood and adolescence but can also be specifically demonstrated for the deterioration of planning ability in old age (Köstering, Leonhart, Stahl, Weiller & Kalle, 2011, in prep.) Similarly, the “Tower of London” enables individual processes involved in planning to be spatially and temporally dissociated from each other on the basis of differences in eye movement patterns caused by the problem structure (Hodgson, Bajwa, Owen & Kennard 2000; Kaller, Rahm, Bolkenius & Unterrainer 2009; Nitschke, Ruh, Kappler, Stahl & Kaller, submitted) and in the brain activation patterns of the areas of the prefrontal cortex that are involved.



TowerTool 2.01 (christoph.kaller@uniklinik-freiburg.de)

Peg Heights: Peg 1 3, Peg 2 2, Peg 3 1; # Balls 3; Tower Type London

	Ball 1 Peg	Ball 1 Pos	Ball 2 Peg	Ball 2 Pos	Ball 3 Peg	Ball 3 Pos	Ball 4 Peg	Ball 4 Pos	Ball 5 Peg	Ball 5 Pos	B&B 2002
Start State	1	1	1	2	1	3					11
Goal State	2	1	1	2	1	1					43

Status

```
--------------------------------------------------------------------------------
17:30:44   LOAD problem set:  1st iso_minmov7.tts
17:30:44   ... set current problem set directory to:  F:\home\kaller\Desktop\SIC
--------------------------------------------------------------------------------
17:30:51   DISPLAY problem from set:  TOL321__Berg[11-43]__Peg[123!!-32
17:30:51   ... NO RESULTS AVAILABLE in current problem set directory
--------------------------------------------------------------------------------
17:31:01   ANALYSIS of problem structure:  TOL321__Berg[11-43]__Peg[321
17:31:01   ... finished at depth of 7 move(s)
17:31:01   ... evaluating OptimalPaths 2 / 2
17:31:01   ... evaluating SubOptimalPaths 6 / 6
17:31:02   ... cleaning up display
17:31:02   ... updating OptimalPaths: 2 / 2
17:31:04   ... updating SubOptimalPaths: 6 / 6
17:31:05   ... updating IsoProblems: 6 / 6
17:31:05   ... DONE
```

Problem Set Control

CD ... ler\Desktop\SICHERUNG\TowerTool\TowerTool Development\examples

```
TOL321__Berg[11-41]__Peg[123!!-
TOL321__Berg[11-43]__Peg[123!!-
TOL321__Berg[11-45]__Peg[123!!-
TOL321__Berg[11-54]__Peg[123!!-
TOL321__Berg[11-55]__Peg[123!!-
TOL321__Berg[12-44]__Peg[12!!3-
TOL321__Berg[12-46]__Peg[12!!3-
TOL321__Berg[12-54]__Peg[12!!3-
TOL321__Berg[12-55]__Peg[12!!3-
TOL321__Berg[13-41]__Peg[12!3!-
```

Results

Structural Problem Parameters

TOL321 Berg[11-43] Peg[321!!-23!1!] Ball[111213-211211]

Minimum Moves: 7 | Suboptimal Paths: 6 | Start: full tower (300) unamb. hierarchy (200)

Optimal Paths: 2 | Detours: 6 | Goal: partial tower (210) comp. amb. hierarchy (110

Search Depth: 3 to 4 | Dead Ends: 0

Optimal Paths

#001 Depth: 3 + 4 ContPaths: 1 / 1

#002 Depth: 4 + 3 ContPaths: 1 / 0

Structural Problem Parameters

TOL321 Berg[13-52] Peg[21!3!-32!!1] Ball[111221-311112]

Minimum Moves: 4 | Suboptimal Paths: 2 | Start: partial tower (210) part. amb. hierarchy (210)

Optimal Paths: 1 | Detours: 1 | Goal: partial tower (210) part. amb. hierarchy (210)

Search Depth: 1 | Dead Ends: 1

Suboptimal Paths

#001 Detour Depth: 2 + 3 ContPaths: 1 / 0

#002 Dead End Depth: 3 + 9 ContPaths: 6 / 5

Figure 2.4 TowerTool is a comprehensive software package that enables systematic analysis of the problem space of disk transfer tasks and the resulting factors that influence the difficulty of individual problems (from Kaller et al. 2011a, where further details can also be found). The latest version of TowerTool can be downloaded at: http://www.uniklinik-freiburg.de/fbi/live/apps/towertasks.html

Taken together the existing studies lay the foundation for a new approach to the operationalization of planning difficulty based on cognitive psychology. A systematic review of the literature on the influence of structural problem parameters on the measurement of different aspects of planning ability can be found in Kaller et al. (2011a), as can a freely accessible computer program for comprehensive problem space analysis of widely used disk transfer tasks such as the “Tower of London”, the “Tower of Hanoi” and variants of them (Figure 2.4).

2.1.4 Development of a "Tower of London" standard problem set

In the past, development of tests based on the "Tower of London" has paid little attention to the influence of problem structure on the measurement of planning ability. In the context of clinical application this therefore raises the question of whether planning tests based on the "Tower of London" (or on the "Tower of Hanoi") in neuropsychological test batteries actually measure the same underlying construct and are therefore interchangeable – as is often implicitly assumed or explicitly stated (see Humes et al. 1997; Welsh, Satterlee-Cartmell & Stine 1999; Zook, Davalos, Delosh & Davis 2004). It is possible that the prevailing heterogeneity of some reported findings on planning impairments in particular neurological and psychiatric conditions (for a summary see Sullivan, Riccio & Castillo 2009) can therefore be ascribed not only to factors such as severity of the disease, co-morbidities, type and status of medication, etc., but also to differences in the sensitivity of the tests used caused by differences in the structural problem parameters[2] of the tasks concerned (for a demonstration based on Parkinson's disease see McKinlay et al. 2008, 2009).

To improve the comparability of future studies Kaller et al. (2011a), on the basis of detailed problem space analysis of the "Tower of London" and a number of preparatory studies of their own, have therefore proposed a standard problem set that would represent an acceptable compromise between, on the one hand, a sufficiently broad range of task difficulty to make the test usable virtually universally and, on the other, acceptable test length and satisfactory psychometric properties (see also Kaller, Unterrainer & Stahl 2012a). Design of the standard problem set focused on two structural problem parameters that exert a key influence on task difficulty. Within the minimum move levels, both parameters were varied systematically in the form of a factorial design in order to optimize the sensitivity of the test to different aspects of planning impairment; other influences were as far as possible kept constant during this process (for further details see Kaller et al. 2011a, 2012a). By this means it was possible to achieve a substantial and almost linear increase in difficulty irrespective of the minimum number of necessary moves. On the basis of this increase in difficulty differentiation over a broad spectrum of planning ability can be achieved (Kaller, Unterrainer & Stahl 2012a). The standard problem set has so far been used successfully in our own studies of healthy samples ranging from children (from the age of 6) to older adults, and in studies of adult patients with craniocerebral trauma, schizophrenia, autism and depression and of children and young people with autism and ADHD (unpublished data/manuscripts in prep.)

The "Tower of London – Freiburg Version" (TOL-F) implemented in the Vienna Test System is based to a large extent on the suggested standard problem set and is a refinement of this set optimized for clinical use (for a detailed description see Kaller et al., in prep.).

2.2 Use of the "Tower of London"

The importance of the frontal lobes for the successful planning of action sequences has long been discussed, largely on the basis of anecdotal accounts of behavioral disorganization in patients with frontal brain lesions (e.g. Harlow 1868; Penfield & Evans 1935; for a historical overview see Owen 1997). Only recently, however, has an increasing amount of experimental evidence been reported in the form of neuropsychological studies of disk transfer tasks such as the "Tower of London" or "Tower of Hanoi" (e.g. Carlin et al. 2000; Glosser & Goodglass 1990; Goel & Grafman 1995; Morris, Miotto, Feigenbaum, Bullock & Polkey 1997; Owen et al. 1990; Shallice 1982).

Since the introduction of the "Tower of London" as a planning test for patients with frontal brain lesions by Shallice (1982) many variants of the test have been used to explore a wide

[2] It should, however, be borne in mind that measurement of planning ability may be influenced not only by differences in the structural parameters of the problems used but also by the way in which the instructions are presented, the provision of information on the minimum number of moves and other factors (Unterrainer et al. 2003).

range of issues in both clinical and non-clinical samples (Berg & Byrd 2002, see Figure 2.2). The “Tower of London” has been widely used not only in lesion studies but also in connection with neurodegenerative and psychiatric illnesses (for an overview see Sullivan et al. 2009).
A literature search carried out in October 2011 in MEDLINE (http://www.ncbi.nlm.nih.gov/pubmed) and PSYCINFO (http://www.apa.org/psycinfo) revealed a total of 330 published journal articles on the “Tower of London”. Of these articles, 208 involved studies to measure restricted planning ability in patients with a wide range of etiologies: frontal brain lesions (n=8), craniocerebral trauma (n=9), Parkinson's disease (n=32), Huntington’s disease (n=4), Alzheimer’s disease (n=9), fronto-temporal dementia (n=4), multiple sclerosis (n=7), substance misuse (n=19), schizophrenia and schizoaffective disorders (n=29), depression (n=9), compulsive disorders (n=8), ADHD (n=17), autism (n=8) and a number of other neurological and psychiatric illnesses.
In summary, diagnostic use of the “Tower of London” is always indicated where impairment of executive functions in general and planning ability in particular is suspected. The “Tower of London” can, however, also be used as an ability test with normal healthy respondents and with specific groups of respondents (e.g. chess players, see Unterrainer et al. 2006, 2011).

2.3 Structure of the “Tower of London – Freiburg Version” (TOL-F)

In terms of the summary of existing taxonomies and explanations of planning and problem-solving in Section 2.1.1 – which makes no claim to be complete – the “Tower of London” can be classed as a well-defined and knowledge-lean planning task that is primarily intended to measure planning ability in the sense of initial planning of action sequences and their consequences. The “Tower of London – Freiburg Version” (TOL-F) implemented in the Vienna Test System is based on the design originally proposed by Shallice (1982) involving three rods of different heights, on which three differently colored balls are placed.

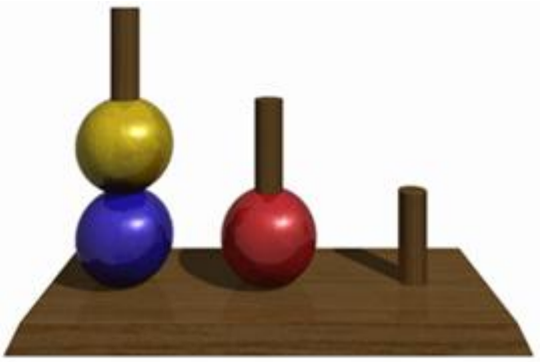


Figure 2.5 Implementation of the “Tower of London” in the Vienna Test System.

Implementation of the “Tower of London” in the Vienna Test System is computerized in the form of a realistic three-dimensional representation of a wooden model of the tower configuration (Figure 2.5). As in the original, the left-hand rod can hold three balls, the centre rod holds two balls and the right-hand one one ball. The ball colors are red, yellow and blue. The green of the original has been replaced here by yellow (see Figure 2.1) in order to make the test usable with respondents affected by red/green color blindness.
For each item that is presented the goal state is always shown in the upper part of the screen and the start state in the lower part. To solve the problem the respondent must convert the start state into the goal state. The minimum number of moves needed to do this is shown on

the left next to the start state. The TOL-F is worked using the computer mouse. The test cannot be worked using a touch screen.

To move a ball in the start state configuration, the respondent must first select it by clicking on it. To indicate the selection that has been made, a white corona appears around the selected ball. By clicking on the desired position, the selected ball can then be moved to this location. Only one ball can be moved at a time. The balls can only be moved onto the rods of the starting configuration shown in the lower part of the screen. Balls that are blocked by balls lying on top of them cannot be selected. Likewise balls cannot be placed on rods that are already full to capacity. Attempts to break these rules when moving and depositing balls are recorded by the computer; this information is available for qualitative evaluation.

There is a time limit on the working of each item. This time limit is 60 seconds and is the same as that used by Shallice (1982). The remaining working time is continuously reported via a bar in the upper right-hand corner of the screen. If the time limit is exceeded, the item that is being worked on is automatically terminated. If the time limit is exceeded in three successive items, the TOL-F is automatically terminated. For patients with cognitive and/or motor retardation there is an option to increase the time limit to three minutes or deactivate it completely[3]. Pauses between items are possible, since the individual problems are not presented automatically but are actively started by the respondent.

2.4 The TOL-F – test forms

The TOL-F consists of two test forms. The standard form provides a detailed assessment of planning ability. The short form discriminates well in the lower ability range and enables quick and economical measurement of performance deficits

Both forms of the TOL-F are designed in the same way. After an instruction and practice phase with two-move problems, simple three-move problems follow and then four-, five-, and six-move problems. The three-move problems represent the familiarization phase that is usual in an ability test in order to avoid distortion of individual test results on account of any remaining difficulties in understanding the task. In both the standard and the short forms of the TOL-F the actual measurement of planning ability is based on the four- to six-move problems that are presented.

The problem sets used in both test forms are derived from the standard problem set proposed by Kaller et al. (2011a, 2012a), refined and optimized for clinical use (for a detailed description see Kaller et al., in prep.).

2.4.1 Standard form of the TOL-F (approx. 16 minutes)

The standard form of the TOL-F consists of 28 items that are presented in the order of an increasing minimum number of moves. There are four three-move problems and eight each of four-move, five-move and six-move problems.

As a result of the range of items used, the standard form is suitable for measuring individual planning ability across a wide ability spectrum. It can therefore be used with a wide range of clinical and non-clinical samples without risk of global floor or ceiling effects or of failure to depict inter-individual differences. However, a wide range of item difficulties is always achieved at the expense of reliability and can only be compensated for by lengthening the test (see bandwith-fidelity dilemma, Cronbach 1984). In clinical settings, however, there are usually tight limits on the time available for administration of a single test. The aim in designing the standard form of the TOL-F was therefore to select a problem set of optimum bandwith with satisfactory reliability (see Section 3.2.1) and at the same time of practical test length (Kaller et al., in prep.).

[3] However, if the time limit is extended or removed, norm scores should be used only with reservation, since these scores were obtained using a one-minute time limit.

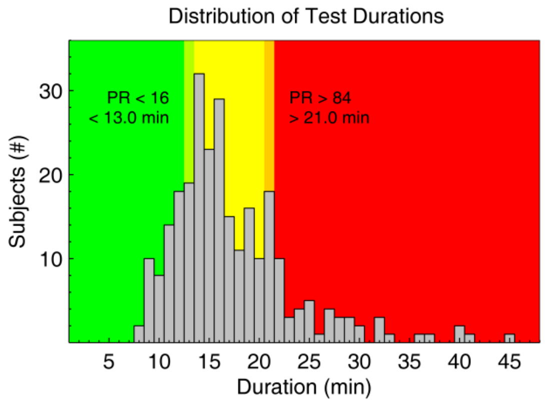


Figure 2.6 Distribution of the duration of the standard form of TOL-F in the norm sample. The color coding differentiates different percentile rank intervals (green, PR < 16, clearly below average; pale green, PR 16-24, below average to average; yellow, PR 25-75, average; orange, PR 76-84, average to above average; red, PR > 84, clearly above average). Abbreviation: PR = percentile rank

In the norm sample (see 4.1) respondents needed on average 17.2 minutes (median: 16 minutes) to complete the standard form of the TOL-F including the instruction and practice phase. Three-quarters of respondents completed the test within 20 minutes and only 16% needed more than 22 minutes (see Figure 2.6). The standard form of the TOL-F is thus of acceptable length for use in clinical investigations. Details of reliability and other test properties are given in chapters 3 and 4.

2.4.2 Short form of the TOL-F (approx. 11 minutes)

In addition to the short form of the TOL-F, the Vienna Test System also contains a short form for quick and economical assessment of performance deficits.

The short form of the TOL-F uses 14 of the 28 items contained in the standard form. These 14 items comprise two three-move and two four-move problems, plus five five-move and five six-move ones. The items were selected using various criteria. Firstly, the short form needed to correlate as closely as possible with the standard form of the TOL-F and also to have maximum reliability. Secondly, the test needed to have high sensitivity (i.e. < 5% false negative classifications), especially in the lower part of the ability range, and acceptable specificity (i.e. < 20% false positive classifications). This means that ideally all respondents who display clearly below average performance on the standard form of the TOL-F (percentile rank <16) should not achieve a better score on the short form. Similarly, the number of respondents who perform averagely on the standard form (percentile rank ≥16) but obtain a below-average score on the short form should be as small as possible.

Items were selected empirically on the basis of a dataset relating to the TOL-F standard form (n=269). After the number of four- to six-move problems in the short form had been set at 12, the properties of all possible problem subsets were considered in depth. On the basis of the above criteria, the final form of the short form was then defined (Kendall's τ = .791; exhaustively estimated split-half reliability, r_{mean} = .613; Cronbach's α = .611; true positive rate, *TPR* = .959; true negative rate, *TNR* = .836). In order to incorporate a familiarization phase (see above), two three-move problems were added.

To validate the item selection, parallel testing with the short and standard forms of the TOL-F was conducted with an independent sample (n=53). Details can be found in chapters 3 and 4. The mean working time, including instructions and practice phase, was 11.8 minutes (median, 11; minimum, 6; maximum 25).

2.4.3 Parallel versions

The Vienna Test System contains three parallel versions of each of the two forms of the TOL-F. The different versions contain systematically varied permutations of the three ball colors (see Berg & Byrd 2002; Unterrainer et al. 2005; Kaller et al. 2011a). The parallel versions thus consist of structurally identical problems of equal item difficulty presented in different visual formats.

In the Vienna Test System the six different parallel versions or color permutations are coded with the letters A to C. Short form A is thus part of standard form A, short form B is part of standard form B and so on. If the short and standard forms of the TOL-F are used at the same time, care should be taken not to use the same color permutations.

2.5 Description of variables

Calculation of the variables is based on all the items that can be solved in four to six moves. By contrast the items at the start of the test that can be solved in three moves serve primarily as practice items.

Variable	**Description**
Planning ability	Number of the four- to six-move items solved in the minimum number of moves.
Non-optimally solved items	Number of items solved in more than the minimum number of moves.
Time limits exceeded	Number of items terminated because the pre-set time limit (1 or 3 minutes per item; can in some circumstances be deactivated) was exceeded.
Reversed decisions	“Undoing” a ball that has already been moved by clicking on it again.
Selecting a blocked ball	Error or infringement of a rule by picking up a ball that is blocked by one above it.
Selecting a blocked rod	Error or infringement of a rule by placing a ball on a rod that is already full.
Selecting an impossible position	Error or infringement of a rule by picking up or placing a ball by clicking outside the area defined by the tower configuration of the start state.
Median planning time	Median planning times, reported separately for four-, five- and six-move problems. Calculation is based only on items in which the goal state was achieved.
Median execution time	Median execution times, reported separately for four-, five- and six-move problems. Calculation is based only on items in which the goal state was achieved.

3 EVALUATION

3.1 Objectivity

3.1.1 Administration objectivity

Test administrator independence exists when the respondent's test behavior, and thus his test score, is independent of variations (either accidental or systematic) in the behavior of the test administrator (Kubinger, 2003).Computerized administration of the TOL-F ensures that all respondents receive the same information, presented in the same way, about the test. These instructions are independent of the test administrator.

Similarly, administration of the test itself is identical for all respondents.

3.1.2 Scoring objectivity

The respondent's answers are registered automatically. Calculation of the test variables and the norm comparison also take place automatically in the TOL-F; a scorer is not involved. Computational errors are thus excluded and a high level of scoring objectivity is ensured.

3.1.3 Interpretation objectivity

Since the TOL-F has been normed, interpretation objectivity is given (Lienert & Raatz, 1994). Interpretation objectivity does, however, also depend on the care with which the guidelines on interpretation given in the chapter "Interpretation of Test Results" are followed.

3.2 Reliability

3.2.1 Standard form of the TOL-F

The reliability of the standard form of the TOL-F was determined using the norm sample (n=269). A detailed description of the sample will be found in Section 4.1.
Overall the reliability of the standard form is satisfactory. Internal consistency was calculated according to Cronbach (1951) in the form of the Alpha coefficient; it was found that α = .7022. Split-half reliability was estimated exhaustively by means of the fully permutated assignment of individual item twins to test halves (see Kaller, Unterrainer & Stahl 2012a); according to the Spearman-Brown formula r_{mean} = .7096 and r_{max} = .7809. As expected, estimation of split-half reliability according to Kristof (1963) produces the same result (r_{mean} = .7096, r_{max} = .7825). In addition, the "greatest lower bound" (glb; Jackson & Agunwamba 1977) was also calculated as an alternative measure of internal consistency. In recent years various authors have proposed this measure as an alternative to the Alpha coefficient (e.g. Sijtsma 2009). The glb of the standard form of the TOL-F was calculated using the R psych package (Revelle, 2011) as .825.
As a result of the manipulation of difficulty that was intended and successfully realized via the minimum number of necessary moves, the standard form of the TOL-F has a large bandwidth (see Section 4.2.1). Accordingly there is no tau-equivalence between the individual items and it can also be assumed that the reliability coefficients quoted represent merely the lower boundary of the true reliability, which is in reality at least as high and probably higher (Bühner 2006).

3.2.2 Short form of the TOL-F

The reliability of the short form of the TOL-F was tested using a sample independent of the norm sample by means of parallel testing with the short and standard forms of the TOL-F. The data was collected in September 2011 in the research laboratory of SCHUHFRIED GmbH in Vienna/Austria. The sample comprises 153 neurologically and psychiatrically healthy individuals (29 or 54.71% of them male) aged between 16 and 73, taken from the normal population. The aim was to achieve a distribution of age and gender that was as nearly uniform as possible. The respondents first completed the short form of the TOL-F and then after a short break the standard form.

The parallel reliability in the form of the estimated correlation between the short and standard forms of the TOL-F is Kendall's τ = .4662. Internal consistency according to Cronbach (1951) is α = .4551 (.4818 according to the Kuder-Richardson formula). Split-half reliability was estimated exhaustively via the fully permutated assignment of individual items to test halves; split-half reliability according to the Spearman-Brown formula is r_{mean} = .4879 und maximal r_{max} = .7371. As expected, estimation of split-half reliability according to Kristof (1963) produces a similar if somewhat higher result (r_{mean} = .5018, r_{max} = .7466). Estimating the internal consistency of the short form from the data of the norm sample yields a value of α = .6345. The value for the glb in this dataset is .747.

By comparison with the reported reliability (Cronbach's α = .25, split-half reliability r = .19; see Humes, Welsh, Retzlaff & Cookson 1997; Schnirman, Welsh & Retzlaff 1998) of the frequently used selection of 12 two- to five-move problems based on Shallice (1982), the short form of the TOL-F, which also consists of 12 items, is clearly superior. Nevertheless, it is recommended that the short form is used only as a coarse screening instrument and only in cases in which completion of the standard form is not possible.

3.3 Validity

Despite criticism of the as yet insufficient consideration given to the influence of structural item parameters on problem difficulty and the heterogeneous findings with regard to some clinical disorders that arises from this and other factors (see Sullivan et al. 2009), the validity of the "Tower of London" paradigm as a test for measuring planning ability is in general supported by extensive literature. In particular, various validation studies are in course of preparation for the TOL-F that is included in the Vienna Test System.

3.4 Scaling and dimensionality of the test

The quality criterion of scaling is met when the empirical behavioral relationships under consideration can be represented exactly by the test scores (Kubinger 2003). As part of the inspection of this psychometric property, some studies of the dimensionality of the TOL-F will be described below.

3.4.1 Description of the Rasch model

The unidimensionality of tests can checked by means of procedures for checking the test model of Rasch (1960). This model assumes that the probability that a particular individual v will solve a particular test item i is specified by a person-specific ability parameter θ_v and an item-specific difficulty parameter ε_I. If these two parameters are known, the probability that person v will solve item i is given by the following equation:

$$P(+\,|\,\theta_v,\varepsilon_i)=\frac{\exp(\theta_v-\varepsilon_i)}{1+\exp(\theta_v-\varepsilon_i)}$$

The validity of the Rasch model also means that the raw score – i.e. the sum of correctly solved items – contains all the information about the ability of the person tested (Rost, 2004).

From this it follows that the pattern of which items were solved or not solved yields no further information about the respondent's ability. A further consequence of the Rasch model is that when it applies the relative items difficulties ε_I are equal for all respondents. The test of this assumption tests an aspect of the test's fairness (see Section 3.9).

The test of model quality was carried out in two stages. Model quality was first tested by means of infit and outfit statistics and by principal component analysis of the Rasch residuals, calculated using the software Winsteps (Linacre, 2007). After this the similarity of the relative item difficulties in different relevant subgroups was investigated. This was done using the R package eRm (Mair, Hatzinger & Maier, 2011). The four-, five- and six-move problems (items 5 – 28) of the standard form of the TOL-F were investigated using the data of the norm sample.

3.4.2 Testing model validity by means of infit and outfit statistics

The standardized infit and outfit statistics test at item level whether the answers actually given by all respondents are to be expected under the assumption of the validity of the Rasch model. The expected value of these statistics is 1. Values below 0.5 indicate that the observed answers can be predicted unexpectedly well; this means that the corresponding items yield very little additional information. By contrast, values over 2.0 indicate that the answers given are highly unexpected under the assumption of the validity of the model; these answers are detrimental to the measurement (Linacre 2007, p. 221f).

Table 3.1: Infit and outfit statistics for the standard form of the TOL-F

Item	Mean-square infit statistic	Mean-square outfit statistic
Item 5	1.01	1.02
Item 6	0.92	0.53
Item 7	0.92	1.05
Item 8	1.01	1.11
Item 9	0.94	0.87
Item 10	0.99	1.11
Item 11	0.92	1.60
Item 12	0.95	0.86
Item 13	0.91	0.94
Item 14	1.01	1.00
Item 15	0.91	0.83
Item 16	0.90	0.93
Item 17	1.02	1.20
Item 18	1.01	1.34
Item 19	1.06	1.18
Item 20	1.21	1.27
Item 21	0.99	0.96
Item 22	0.92	0.90
Item 23	0.97	0.94
Item 24	1.03	1.14
Item 25	0.95	0.94
Item 26	1.03	1.07
Item 27	0.99	0.98
Item 28	1.16	1.20

The resulting infit and outfit statistics for the TOL-F are shown in Table 3.1. The infit statistics are particularly sensitive to unexpected response behavior in items that correspond with the testee's ability level. The outfit statistics, on the other hand, are sensitive to unexpected response behavior in items that are too easy or too difficult for the respondent. Overall the Rasch model is shown to fit sufficiently well at individual item level.

3.4.3 Testing model validity by means of principal component analysis of the Rasch residuals

In a second step principal component analysis was carried out on the Rasch residuals (i.e. the differences between 1 and the predicted probability of the answer in question). This analysis serves to identify any systematic model deviations which could indicate that a second ability dimension is involved. In this form of model test the eigenvalues are first calculated. On the basis of the estimated person and item parameters, datasets are then simulated that fit the Rasch model perfectly. If the results obtained for this simulated data are similar to those for the original dataset, this is evidence of the validity of the Rasch model.
The relevant principal component analysis yielded eigenvalues of 1.7 for the first and second factors. The subsequent data simulation confirmed that such results are to be expected if the Rasch model is valid. Overall these findings thus indicate that the TOL-F meets the criterion of unidimensionality sufficiently well.

3.4.4 Testing person homogeneity

A third step tested the assumption of person homogeneity, which means that the relative item difficulties remain the same in different subsamples. The statistical procedure used for this purpose is the Likelihood Quotient Test of Andersen (1973). Results for the splitting criteria of age (divided according to median age, i.e. individuals up to the age of 45 and individuals 46 and over), gender, education (individuals with a VTS educational level of up to 3 and individuals with a VTS educational level of 4 or higher) and test performance (divided according to the median) are shown in Table 3.2.

Table 3.2 Goodness of fit for the variable Planning ability

Splitting criterion	χ^2	***df***	***p***
Age	41.126	23	0.011
Gender	29.793	23	0.155
Education	57.818	23	<0.01
Test performance	34.943	22	0.039

Overall it was found that the relative item difficulties remain comparable among the subsamples for almost all the splits undertaken. Only for the splitting criterion of education is there a result that is significant at the 1% level. The Likelihood Quotient Test is particularly sensitive to infringements of person homogeneity. Further item bias analysis yields sufficiently high correlations (Spearman's ρ > .94), indicating that there are only slight differences in relative item difficulty in the various subsamples. A corresponding comparison of the parameter estimates of item difficulty for splits according to test performance, age, gender and education is shown in Figure 3.1.

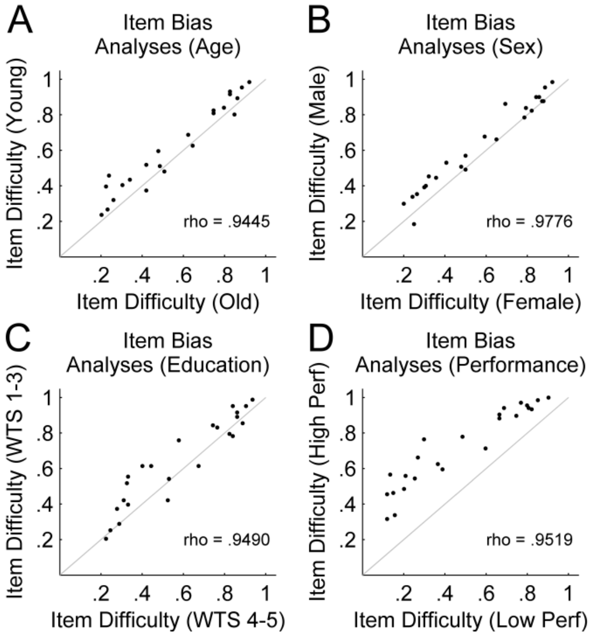


Figure 3.1 Comparison of the difficulty parameters for individuals (A) of low and high age, (B) male and female gender, (C) with low and high education, and (D) with low and high test performance.

3.5 Economy

Being a computerized test, the TOL-F is very economical to administer and score. The administrator’s time is saved because the instructions at the beginning of the test are computerized, relieving him of the need to provide time-consuming verbal explanations. Because the test results are calculated automatically, the time needed for manual calculation of raw and norm scores is also saved.

3.6 Usefulness

The quality criterion of usefulness is met if a test (1) measures a relevant trait and (2) this trait cannot be measured by other tests that meet all the other quality criteria to at least the same extent. (Kubinger 2003).
Successful and autonomous coping with everyday life requires a high degree of adaptability, especially in situations that deviate from the usual routine and for which it is not possible to fall back on existing action schemata. By measuring the executive functions for action planning and monitoring that are needed in such situations, the TOL-F thus measures a trait that is highly relevant and to which little attention is paid as an independent construct in other tests of executive functions (see Unterrainer et al. 2003).

3.7 Reasonableness

In order to meet the quality criterion of reasonableness, tests must be so constructed that the respondent is not overstretched physically and is not put under psychological stress either emotionally or in terms of energy and motivation. This applies at all times, but needs in particular to be borne in mind in relation to the diagnostic context in which the test is being used (e.g. Kubinger 2003). The TOL-F enables the duration and difficulty of testing to be adapted flexibly to the particular respondent. For example, with severely handicapped patients the short form of the TOL-F can be used alone for orientation purposes. By contrast, with able patients the standard form of the TOL-F can be used; this enables differentiated assessment of planning ability.

3.8 Resistance to faking

A test that meets the meets the quality criterion of resistance to faking is one that can prevent a respondent answering questions in a manner deliberately intended to influence or control his test score (see Kubinger 2003). Since the TOL-F is an ability test, faking in the sense of “faking good” is not possible. “Faking bad” can be prevented by creating a test setting in which the respondent feels at ease and by remaining observant and carrying out plausibility checks during the testing session. Common to all the items is the fact that respondents may fail to find the best solution if they initiate action prematurely – i.e. without comprehensive planning of the solution. For this reason the importance of planning the solution is particularly stressed in the instructions. In addition, mean reaction times are given so that the corresponding parameters can be monitored.

3.9 Fairness

If tests are to meet the quality criterion of fairness, they must not systematically discriminate against particular groups of respondents on the grounds of their sociocultural background (Kubinger, 2003). The fairness of the TOL-F is given, since divided norm samples exist for subgroups for which relevant mean differences were found (see also Sections 3.4.4 and 4.2.1).

4 NORMING

The norms were obtained by calculating the mean percentile rank *PR(X)* for each raw score *X* according to the formula (from Lienert & Raatz, 1998):

$$PR_x = 100 \cdot \frac{cum\ f_x - f_x/2}{N}$$

cum fx corresponds to the number of respondents who have achieved the raw score *X* or a lower score, *fx* is the number of respondents with the raw score *X* and *N* is the size of the sample.

4.1 Norm sample

The norm data for the TOL-F was collected between April and September 2011 in the research laboratory of SCHUHFRIED GmbH in Vienna/Austria. The sample comprises 269 individuals (129 or 48% of them male) aged between 16 and 84, taken from the normal population. Individuals were only included if they had no previous neurological or psychiatric disease and were not taking any drugs that act on the central nervous system. Respondents' educational level was assessed via the entry made in the Vienna Test System. The distribution of age, gender and educational level within the norm sample is shown in Figure 4.1 A and B.

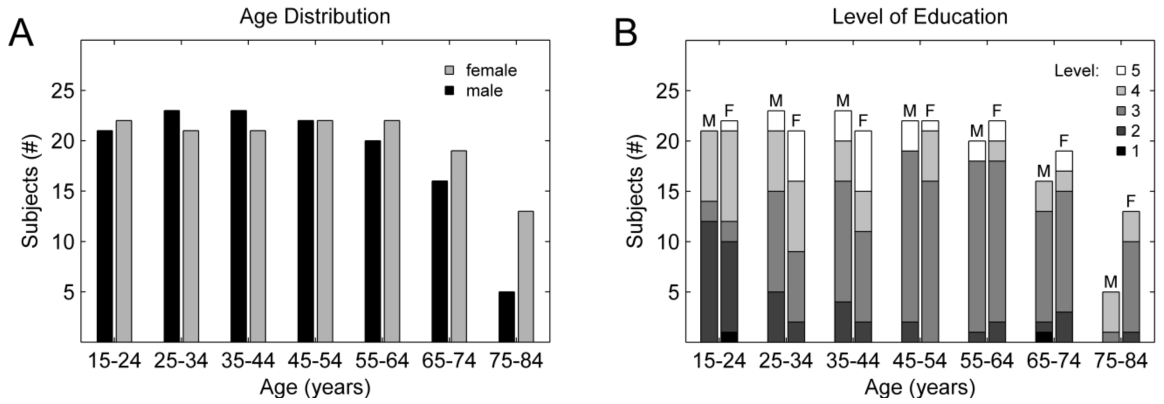


Figure 4.1 (A). Distribution of age in the norm sample in relation to gender. Respondents are divided into seven 10-year cohorts from 15;0-24;11 years to 75;0-84;11 years. (B) Distribution of educational level within the norm sample in relation to age and gender (VTS educational level: Level 1, compulsory schooling not completed; 2, compulsory schooling or basic secondary school; 3, technical school or vocational training; 4, upper secondary school with leaving examination at university entrance level; 5, university).

A key aim in collecting the norm data was to have an age distribution that was as uniform as possible for both genders, in order to ensure that the quality of norming was independent of both age and gender. This has been achieved up to age of 74 (see Figure 4.1 A). Norm data in the age range above 75 years is only conditionally usable in the current version of the TOL-F. There are, however, plans to continue the collection of norm data in the upper age range and to make this available as soon as possible.

4.2 Test forms

4.2.1 Standard form of the TOL-F

Measuring individuals' planning ability in the context of clinical and/or scientific investigations requires a set of problems that can be used for testing across a wide range of ability levels and that discriminates between respondents of differing ability. In developing the standard form of the TOL-F the emphasis was therefore on ensuring a continuously rising level of difficulty across the minimum number of necessary moves (see 2.4.1; see also Kaller et al. 2011a, 2012a).

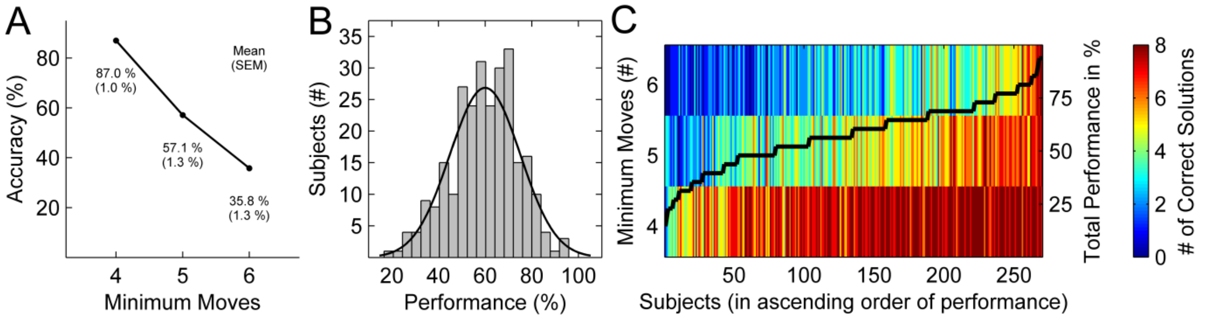


Figure 4.2 (A) Number of correctly solved problems in relation to the minimum number of necessary moves. (B) Distribution of performance within the norm sample. (C) Performance of individual respondents in relation to item difficulty in the sense of the minimum number of necessary moves (left ordinate). The respondents are arranged along the abscissa in ascending order of planning ability. The black line shows the number of items worked correctly by each individual (right ordinate). The color coding indicates the number of correctly solved problems for the minimum number of necessary moves.

Analysis of the norm sample shows that the standard form of the TOL-F includes a wide range of item difficulties related to the minimum number of necessary moves (Figure 4.2 A; ANOVA with measurement repetition; significant main effect for number of moves, $F_{(2,536)}$ = 696.6, p < .001, partial η^2 = .722). Furthermore, individual comparisons made by contrasting four-move vs. five-move problems ($F_{(1,268)}$ = 517.6, p < .001, partial η^2 = .659) and five-move vs. six-move problems ($F_{(1,268)}$ = 222.3, p < .001, partial η^2 = .453) confirm that difficulty increases continuously across the minimum number of necessary moves. For a range of medium item difficulty from 87.0% (four-move problems) to 35.8% (six-move problems; see Figure 4.2 A) it can therefore be assumed that the standard form of the TOL-F is almost universally usable for differential assessment of planning ability..

Despite minor deviations (Kolmogorov-Smirnov Test, χ^2 = 1.41, p = .038), performance on the standard form of the TOL-F also has an approximately normal distribution (Figure 4.2 B); within the norm sample performance spans a range from 16.7% (4 out of a maximum of 24 problems correctly solved) to 95.8% (23 problems).The mean performance is 59.9% (14.4 problems) with a standard deviation of 15.1% (3.6 problems). Figure 4.2 C shows that the operationalization of item difficulty used in the TOL-F also produces results at individual level that are in themselves consistent. In other words respondents who perform less well on simple problems also perform less well on more complex items, while respondents who perform well on more complex problems also solve simpler items reliably.

To ensure the psychometric property of *test fairness* additional analysis was carried out to confirm that performance on the test is independent of the variables of age (10-year cohorts, see Figure 4.1) and gender (ANOVA; significant main effect for age, $F_{(6,255)}$ = 3.5, p = .002, partial η^2 = .076; significant main effect for gender, $F_{(1,255)}$ = 6.9, p = .009, partial η^2 = .026; interaction effect non-significant, $F_{(6,255)}$ = 1.3, p = .264, partial η^2 = .029). This shows that both age and gender influence planning ability as measured by the standard form of the TOL-F (Figure 4.3).

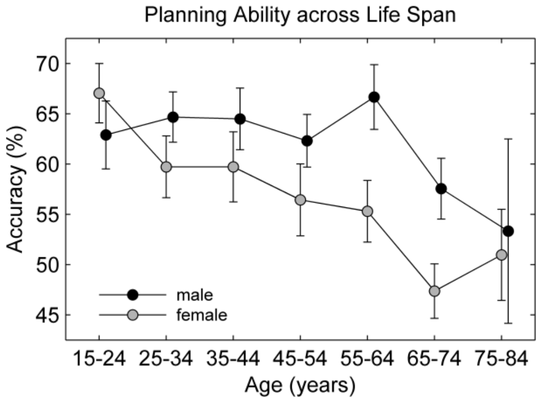


Figure 4.3 Planning ability in relation to age and gender.

For age the mean performance difference between the youngest and the oldest cohorts is 13.3% (3.2 items); for gender the mean performance difference is 5.8% (1.4 items). Converted into effect strengths, age therefore appears to have a stronger effect on performance (d=.843) than gender, for which the effect is only small to moderate (d=.385). In consequence the present norming of the standard form of the TOL-F takes account of age but not of gender, since norms should always be based on sufficiently large (sub)samples in order to provide the most accurate estimate possible of the population in question (Bühner 2006). However, there are plans to provide gender-specific norms in the near future as part of the ongoing collection of norm data (see 4.1). Figure 4.4 provides a summary of the resulting distribution of cut-off scores for the age-appropriate delimitation of below-average and above-average planning ability. Detailed norm tables can be consulted in the Vienna Test System via the Norm Table Explorer (accessed via “Extras” on the menu).

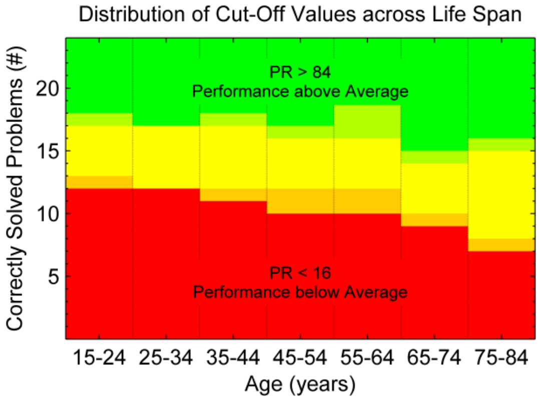


Figure 4.4 Distribution of cut-off norms for the number of correctly solved problems in the standard form of the TOL-F in relation to age (10-year cohorts). The color coding differentiates different percentile rank intervals (red, PR < 16, clearly below average; orange, PR 16-24, below average to average; yellow, PR 25-75, average; pale green, PR 76-84, average to above average; green, PR > 84, clearly above average). Abbreviation: PR = percentile rank

4.2.2 Short form of the TOL-F

The properties of the short form with regard to the classification of clearly below average performance was tested empirically using the parallel test sample (n=53), members of which had completed both the short form and the standard form of the TOL-F (see 3.2.2). Performance on the standard form served as the criterion, while performance on the short form was the predictor. The accuracy of the short form was found to be 81.13%; sensitivity was 62.5% and specificity 84.44%. Because of the small size of the sample and hence the small number of respondents who obtained clearly below-average results (n=8), it is likely that the actual sensitivity of the short form has been underestimated. This will be tested when further data on parallel test reliability is collected. For the time being the norm tables for the short form of the test are based on data collected using the standard form of the TOL-F. There are plans to create a separate norm sample for the short form of the TOL-F in the near future. For the time being it is recommended that the short form is used only as a coarse screening instrument and only in cases in which completion of the standard form is not possible.

5 TEST ADMINISTRATION

5.1 Instruction phase

The instructions at the start of the test can be followed independently by the respondent on his screen; the test administrator is not required to provide any further explanation. Both the standard form and the short form of the TOL-F include standardized instructions with a practice phase. When working with patients it is nevertheless advisable for the administrator to assist during the instruction phase and during testing, too, to check at least from time to time that subjects are working in accordance with the instructions. During the instruction and practice phase feedback is provided if the respondent does not comply with the instructions or if his behavior indicates that the instructions have not been understood. In this case the instruction and practice phase must be repeated.

5.2 Test administration

The instruction and practice phase is immediately followed by the test phase. This takes different forms in the different versions of the TOL-F. In the standard form the test phase lasts about 16 minutes; in the short form it takes about 11 minutes (see also Section 2.4).
Both forms of the TOL-F commence with some instructions. The manner of working the test is then explaining using two-move problems as an example and the respondent's understanding of the task is checked. For familiarization purposes and to consolidate comprehension of the task, either two or four three-move problems are then presented before the actual measuring of planning ability – using four- to six-move problems – begins. For further details see also Section 2.3.

6 INTERPRETATION OF TEST RESULTS

Interpretation is based on the percentile rank scores of the test results for the individual variables of the TOL-F test forms. Interpretation of the content of the main and secondary variables of the TOL-F is described after the general notes on interpretation.

6.1 General notes on interpretation

In general a percentile rank of <16 can be interpreted as a below-average score on the corresponding variable. A percentile rank between 16 and 24 can be regarded as representing a below-average to average level of the corresponding variable. A percentile rank of 25 to 75 can be regarded as average. Percentile ranks between 76 and 84 are average to above average and a percentile rank larger than 84 is clearly above average.
The norm scores always relate to the particular reference population used. For the standard form of the TOL-F age-appropriate norm scores are provided. There are plans to add gender-specific norms in the near future as part of ongoing expansion of the norm sample.
The norm scores derived from the norm sample are applicable to all the parallel versions. It should, however, be borne in mind that planning and execution times reduce non-specifically upon repeated administration of the test (including administration of parallel versions) (see Kaller et al. 2011a, Supplementary Analyses). The measurement of planning ability is affected only very slightly or not at all by repetition of the test. In the event of frequent repetition within a short time, however, overlearning of the items cannot be excluded.
The norm data was collected with a time limit of one minute per item. If the time limit is extended to three minutes or removed entirely the norm scores can only be used with reservation.

6.2 Interpretation of the main and subsidiary variables

6.2.1 Standard form of the TOL-F

Planning ability

This is the main variable that is key to interpretation of an individual’s results on the TOL-F. It describes a person’s ability to plan ahead in a specified context using clearly defined rules and thus to arrive at a correct solution in the optimum way.

Non-optimally solved items

Subsidiary variable; it includes the items for which the time limit was exceeded and so is complementary to *planning ability*.

Reversed decisions

Subsidiary variable that captures corrections that the respondent decides to make while working the test – i.e. occasions on which a ball that has been picked up is returned to its original position. If the number is raised, especially in combination with significantly reduced planning times, this may be an indication of inadequate planning behavior and a tendency to give way prematurely to impulses without first considering the consequences.

Selection of a blocked ball or rod or of an impossible position.

Subsidiary variable that captures the tendency to avoid or ignore rules. However, while a raised number on this variable may indicate non-compliance with the rules, it may also be caused by motor inaccuracies when picking up or putting down the balls.

Median planning time

Subsidiary variable that measures planning time separately for four-, five- and six-move problems. Significantly lowered or raised planning times can provide further information about the causes of planning abilities that differ from the norm, for example in patients with impairments of impulse control or working memory. If testing is repeated, though, the norm scores should be interpreted only with reservations.

Median execution time

Subsidiary variable that measures execution time separately for four-, five- and six-move problems. If testing is repeated the norm scores should be interpreted only with reservations.

Time limits exceeded

The test protocol (displayable via the Vienna Test System's display options) shows for which items the time limit was exceeded. When the time limit is exceeded planned times are usually significantly raised; in some cases execution is not commenced within the specified time limit. On the other hand, where exceeding the time limit is combined with significantly shortened planning times, this indicates that the respondent commenced execution prematurely and without fully planning a solution; as a result he has started to go down a sub-optimal solution path and then made another attempt to find an optimum solution. However, a raised number of occasions on which the time limit is exceeded may also be a sign of cognitive and/or motor retardation. In this case it is recommended that the test is administered with the time limit increased to three minutes or removed entirely; in this situation the norm data can be used only with reservation.

6.2.2 Short form of the TOL-F

Evaluation of the short form of the TOL-F involves the same main and subsidiary variables as in the standard form. The comments on interpretation of the main variables in the standard form thus apply also to interpretation of the short form. However, it should be borne in mind that the norm tables of the short form are for the time being based on the data of the standard form. For this reason, therefore, a cut-off score of 4 for the main variable "planning ability" is given as an additional aid to interpretation. During administration of the items of the short form in the course of norming of the standard form, this score was reached or exceeded by 93% of the sample; by contrast, when the short form was administered on its own to a sample of 53 people, this score was reached or exceeded by 77% of respondents. Scores below this cut-off point can therefore be interpreted as an indicator of possible deficits in planning ability. This can be clarified by subsequently administering the standard form.

6.3 Case study - schizophrenia

Cognitive functions can be impaired long-term or permanently in patients with schizophrenia or schizoaffective disorders. The effect of these cognitive impairments on the patient's ability to return to work, continue in education or training and integrate into family life is often more long-lasting than the acute or residual psychopathology. A comprehensive neuropsychologically oriented assessment is therefore a good predictor of the capacity to work.

Mr P. is a paranoid schizophrenic, although currently without acute symptoms. He is taking part in a reintegration program designed to enable him to return to his job as a commercial administrator. In the theoretical training sessions it becomes clear that he is good at accessing his existing knowledge. When carrying out practical tasks with the training company, however, it is noticeable that he seems to lose track of the bigger picture; he tends to spend a lot of time tackling things randomly by trial and error. He then underwent a neuropsychological investigation that yielded the following results:

Table 6.1 Results of neuropsychological testing of Mr P.

Ability	Test	Result
Reaction speed	WTS: WAFA	average
Attention	WTS: WAFG	slightly below average
Memory span	WTS: Corsi	average
Learning ability	CVLT	learning – good average, free recall – good average, long-term recall – good average, recognition - average
Working memory	WTS: N-Back verbal	slightly below average
Everyday planning	BADS: Zoo	delayed, at second attempt without errors
Planning ability	WTS: TOL-F	highly conspicuous, many reversed decisions
Response suppression	WTS: INHIB - GoNoGo	raised error rate combined with above averagely fast responses

From the neuropsychological profile it is clear that there is an accumulation of deficits: slight problems with division of attention and working memory, that seem compensatable when considered in isolation, lead to Mr P. being overchallenged when faced with more complex problem-solving tasks that require integration of monitoring, attention and flexibility skills. This is manifested in the neuropsychological testing as very poor performance on the TOL-F. At work this leads to uncertainty in making decisions and over-hasty action. He is not good at suppressing behavioral impulses and/or considering their usefulness. The patient loses oversight of the issue and is sidetracked by details. These results indicate that training in planning and problem-solving skills is called for.

7 REFERENCES

Andersen, E.B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38, 123-140.

Anderson, J.R. (2005). *Cognitive psychology and its implications*. New York, NY: Worth Publishers.

Anzai, Y., & Simon, H.A. (1979). The theory of learning by doing. *Psychological Review*, 86, 124-40.

Berg, W.K., & Byrd, D. (2002). The Tower of London spatial problem-solving task: Enhancing clinical and research implementation. *J Clin Exp Neuropsychol*, 24 (5), 586-604.

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.

Borys, S.V., Spitz, H.H., & Dorans, B.A. (1982). Tower of Hanoi performance of retarded young adults and non-retarded children as a function of solution length and goal state. *J Exp Child Psychol*, 33 (1), 87-110.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion*. Munich: Pearson.

Burgess, P., Simons, J., Coates, L., & Channon, S. (2005). The search for specific planning processes. In R. Morris & G. Ward (Eds.), *The cognitive psychology of planning* (p. 199-227). Hove: Psychology Press.

Carlin, D., Bonerba, J., Phipps, M., Alexander, G., Shapiro, M., & Grafman, J. (2000). Planning impairments in frontal lobe dementia and frontal lobe lesion patients. *Neuropsychologia*, 38 (5), 655-65.

Claus, N. (1884). La Tour d’Hanoï, Jeu de calcul. *Science et Nature*, 1 (8), 127-8.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.

Cronbach, L.J. (1984). *Essentials of psychological testing* (4th ed.), New York: Harper & Row.

Davies, S.P. (2003). Initial and concurrent planning in solutions to well-structured problems. *Q J Exp Psychol* A, 56 (7), 1147-64.

Davies, S.P. (2005). Planning and problem solving in well-defined domains. In R. Morris & G. Ward (Eds.), *The cognitive psychology of planning* (p. 35-52). Hove: Psychology Press.

Glosser, G., & Goodglass, H. (1990). Disorders in executive control functions among aphasic and other brain-damaged patients. *J Clin Exp Neuropsychol*, 12 (4), 485-501.

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

Goel, V., & Grafman, J. (1995). Are the frontal lobes implicated in „planning“ functions? Interpreting data from the Tower of Hanoi. *Neuropsychologia*, 33 (5), 623-42.

Goel, V., & Grafman, J. (2000). Role of the right prefrontal cortex in ill-structured planning. *Cognitive Neuropsychology*, 17 (5), 415-36.

Harlow, J.M. (1868). Recovery from the passage of an iron bar through the head. *Boston Medical Surgery Journal*, 2, 327-46.

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 (17), 2934-47.

Hodgson, T.L., Bajwa, A., Owen, A.M., & Kennard, C. (2000). The strategic control of gaze direction in the Tower-of-London task. *J Cogn Neurosci*, 12 (5), 894-907.

Humes, G.E., Welsh, M.C., Retzlaff, P., & Cookson, N. (1997). Towers of London and Hanoi: Reliability and validity of two executive function tasks. *Assessment*, 4 (3), 249-257.

Jackson, E.H., & Agunwamba, C.C. (1977). Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items: I. Algebraic lower bounds. *Psychometrika*, 42, 567-578.

Kafer, K.L., & Hunter, M. (1997). On testing the face validity of planning/problem-solving tasks in a normal population. *J Int Neuropsychol Soc*, 3 (2), 108-19.

Kaller, C.P. (2010). *Der Einfluss struktureller Problemparameter auf Planungsleistungen*. University of Freiburg: unpublished dissertation.

Kaller, C.P., Heinze, K., Frenkel, A., Läppchen, C.H., Unterrainer, J.M., Weiller, C., Lange, R., & Rahm, B. (2012b). Differential impact of continuous theta-burst stimulation over left and right DLPFC on planning. Human Brain Mapping. Advanced online publication.

Kaller, C.P., Rahm, B., Bolkenius, K., & Unterrainer, J.M. (2009). Eye movements and visuospatial problem solving: Identifying separable phases of complex cognition. *Psychophysiology*, 46 , 818-30.

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

Kaller, C.P., Rahm, B., Spreer, J., Mader, I., & Unterrainer, J.M. (2008). Thinking around the corner: The development of planning abilities. *Brain Cogn*, 67 (3), 360-70.

Kaller, C.P., Rahm, B., Spreer, J., Weiller, C., & Unterrainer, J.M. (2011b). Dissociable contributions of left and right dorsolateral prefrontal cortex in planning. *Cereb Cortex*, 21, 307-17.

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. *Brain Res Cogn Brain Res*, 20 (3), 462-72.

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

Klahr, D., & Robinson, M. (1981). Formal assessment of problem-solving and planning processes in preschool children. *Cognitive Psychology*, 13, 113-48.

Knoblich, G., & Öllinger, M. (2008). Problemlösen und logisches Schließen. In J. Müsseler (Ed.), *Allgemeine Psychologie* (p. 552-98). Heidelberg: Spektrum.

Köstering, L., Leonhart. R., Stahl, C., Weiller, C., Kaller, C.P. (2011). Altersassoziierte Veränderungen von Planungsfähigkeiten in Abhängigkeit spezifischer Planungsanforderungen. Eine Studie mit dem Turm von London. Poster presented at the DGP conference *“Psychology and the brain”,* 23 – 26 June 2011, Heidelberg. manuscript in prep.

Köstering, L., McKinlay, A., Stahl, C., & Kaller, C.P. (submitted). Differential patterns of planning impairments in Parkinson's disease and sub-clinical signs of dementia? A latent-class model-based approach.

Kotovsky, K., Hayes, J.R., & Simon, H.A. (1985). Why are some problems hard? Evidence from Tower of Hanoi. *Cognitive Psychology*, 17, 248-294.

Kristof, W. (1963). Die Verteilung aufgewerteter Zuverlässigkeitskoeffizienten auf der Grundlage von Testhälften. *Archiv für die gesamte Psychologie*, 115, 230-240.

Kubinger, K.D. (2003). Gütekriterien. In K.D. Kubinger & R.S. Jäger (Eds.),*Schlüsselbegriffe der Psychologischen Diagnostik* (pp. 195-204). Weinheim: Beltz.

Lienert, G.A. & Raatz, U. (1998). *Testaufbau und Testanalyse*. Weinheim: PVU.

Linacre, J.M. (2007). Winsteps (Version 3.61.2) [Computer Software]. Chicago: Winsteps.com.

Mair, P., Hatzinger, R., & Maier, M. (2011). eRm: Extended Rasch Modeling. R package version 0.14-0. http://CRAN.R-project.org/package=eRm

McKinlay, A., Grace, R.C., Kaller, C.P., Dalrymple-Alford, J.C., Anderson, T.J., Fink, J., et al. (2009). Assessing cognitive impairment in Parkinson’s disease: A comparison of two tower tasks. *Applied Neuropsychology*, 16 (3), 177-85.

McKinlay, A., Kaller, C.P., Grace, R.C., Dalrymple-Alford, J.C., Anderson, T.J., Fink, J., et al. (2008). Planning in Parkinson’s disease: A matter of problem structure? *Neuropsychologia*, 46 (1), 384-9.

Morris, R.G., Miotto, E.C., Feigenbaum, J.D., Bullock, P., & Polkey, C.E. (1997). The effect of goal-subgoal conflict on planning ability after frontal- and temporal-lobe lesions in humans. *Neuropsychologia*, 35 (8), 1147-57.

Newell, A., & Simon, H. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nitschke, K., Ruh, N., Kappler, S., Stahl, C., & Kaller, C.P. (submitted). Dissociable stages of problem solving (I): Temporal characteristics revealed by eye-movement analyses.

Ormerod, T.C. (2005). Planning and ill-defined problems. In R. Morris & G. Ward (Eds.), *The cognitive psychology of planning* (p. 53-70). Hove: Psychology Press.

Owen, A.M. (1997). Cognitive planning in humans: Neuropsychological, neuroanatomical and neuropharmacological perspectives. *Progress in Neurobiology*, 53 , 431-50.

Owen, A.M., Downes, J.J., Sahakian, B.J., Polkey, C.E., & Robbins, T.W. (1990). Planning and spatial working memory following frontal lobe lesions in man. *Neuropsychologia*, 28 (10), 1021-34.

Penfield, W., & Evans, J. (1935). The frontal lobe in man: A clinical study of maximum removals. *Brain*, 58, 115-33.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Revelle, W. (2011). psych: Procedures for Personality and Psychological Research Northwestern University, Evanston, http://personality-project.org/r/psych.manual.pdf, 1.0-97

Rost, J. (2004). *Lehrbuch Testtheorie, Testkonstruktion*. Bern: Huber.

Ruh, N., Rahm, B., Unterrainer, J.M., Weiller, C., & Kaller, C.P. (eingereicht). Dissociable stages of problem solving (II): First evidence for process-contingent temporal order of activation in dorsolateral prefrontal cortex.

Schnirman, G.M., Welsh, M.C., & Retzlaff, P.D. (1998). Development of the Tower of London-Revised. *Assessment*, 5 (4), 355-60.

Shallice, T. (1982). Specific impairments of planning. *Phil Trans R Soc London*, 298, 199-209.

Sijtsma, K. (2009). On the use, misuse, and very limited usefulness of Cronbach's alpha. Psychometrika, 74, 107-120.

Simon, H.A. (1975). The functional equivalence of problem solving skills. *Cognitive Psychology*, 7, 268-88.

Spitz, H.H., Webster, N.A., & Borys, S.V. (1982). Further studies of the Tower of Hanoi problem-solving performance of retarded young adults and nonretarded children. *Developmental Psychology*, 18, 922-30.

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.

Unterrainer, J.M., Kaller, C.P., Leonhart, R., & Rahm, B. (2011). Revising superior planning performance in chess players: The impact of time restriction and motivation aspects. *Am J Psychol*, 124 (2), 213-25.

Unterrainer, J.M., Kaller, C.P., Halsband, U., & Rahm, B. (2006). Planning abilities and chess: A comparison of chess and non-chess players on the Tower of London task. *Br J Psychol*, 97 (3), 299-311.

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. *Brain Res Cogn Brain Res*, 23 (2-3), 418-28.

Unterrainer, J.M., Rahm, B., Leonhart, R., Ruff, C.C., & Halsband, U. (2003). The Tower of London: The impact of instructions, cueing, and learning on planning abilities. *Brain Res Cogn Brain Res*, 17 (3), 675-83.

Ward, G., & Allport, A. (1997). Planning and problem-solving using the five-disc Tower of London task. *Q J Exp Psychol*, 50A, 49-78.

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

Welsh, M.C., Satterlee-Cartmell, T., & Stine, M. (1999). Towers of Hanoi and London: Contribution of working memory and inhibition to performance. *Brain Cogn*, 41 (2), 231-42.

Zhang, J., & Norman, D.A. (1994). Representations in distributed cognitive tasks. *Cognitive Science*, 18, 87-122.

Zook, N.A., Davalos, D.B., Delosh, E.L., & Davis, H.P. (2004). Working memory, inhibition, and fluid intelligence as predictors of performance on Tower of Hanoi and London tasks. Brain Cogn, 56 (3), 286-92.