Journal of the International Neuropsychological Society (1997), **3**, 108–119.

On testing the face validity of planning/problem-solving tasks in a normal population

K. L. KAFER AND M. HUNTER
Department of Psychology, University of Newcastle, NSW 2308, Australia

(RECEIVED September 22, 1995; ACCEPTED March 22, 1996)

Abstract

Clinically, tests of executive functions tend to be chosen on face validity. If such tests are to be used to evaluate a clinical population, their ability to measure executive functions should be reliably demonstrated in a normal population. In order to investigate the reliability of such tests, a sample of 130 normal adults (74 women, 56 men) ages 17 to 55 years were administered 4 tests purporting to measure planning/problem-solving: the Tower of London Test, the Six Element Test, the Twenty Questions Test, and the Rey Complex Figure Test. A structural equation modeling approach provided by the LISREL 8 program was used to evaluate three models hypothesized to explain the relationship among the test variables and the latent construct of planning/problem-solving. An adequate model was unable to be estimated, thus raising questions about the meaning of the latent construct planning/problem-solving and the psychometric structure of the Tower of London Test. (*JINS*, 1997, *3*, 108–119.)

Keywords: Executive functions, Planning/problem-solving, Test validity

INTRODUCTION

The notion that the disorganization of behavior that follows a frontal lobe lesion represents a deficit of the ability to plan, program and verify one's activity (Luria, 1973) has been applied in explaining the difficulties exhibited by patients with frontal lobe injuries (Vilkki & Holst, 1991). Also, it has been suggested that such patients may perform normally on neuropsychological tests yet display marked deficits in the performance of everyday life activities (Shallice & Burgess, 1991).

In 1986 Baddeley coined the term "dysexecutive syndrome" for a class of disorders involving higher level cognitive dysfunction (Morris et al., 1990). Executive functioning, as described by Lezak, involves "those capabilities that enable a person to engage in independent, purposive, self-serving behavior successfully" (Lezak, 1983, p. 28). Specific executive functions include planning, checking, monitoring, testing, evaluating, and revising. They entail the ability to mobilize, allocate, and coordinate cognitive resources in order to solve problems, achieve goals and manage daily activities (Wong, 1985; Zec et al., 1992). The importance of these functions to everyday living means that the rehabilitation of executive functioning is crucial for a patient's postinjury reintegration into normal social and community living (Sohlberg & Mateer, 1989). It follows that it would be useful to develop valid and reliable measures with predictive power to measure executive functions.

Unfortunately, measures of these functions have not been able to differentiate between various aspects of executive functioning in a clear and reliable manner (Vilkki & Holst, 1991). Clinicians have tended to choose tests for their face validity, rather than for their psychometric properties (Pusakulich, 1992), and some tests purporting to measure executive functions lack adequate (or even any) normative or control data. If such tests are to be used to evaluate clinical populations reliably, then their ability to measure executive functions should be demonstrated in normal populations.

Therefore, the present study attempts to examine the psychometric properties and relationships among four tests purporting to measure the underlying construct of *planning/problem-solving*. The tests are: (1) the Tower of London Test (Shallice, 1982), (2) the Six Element Test (Shallice & Burgess, 1991), (3) the Twenty Questions Test (Mosher & Hornsby, 1966), and (4) the Complex Figure Test (Rey, 1941).

The Tower of London Test was designed by Shallice (1982) to investigate the planning abilities of patients with frontal lobe

Reprint requests to: Michael Hunter, Department of Psychology, The University of Newcastle, NSW 2308, Australia.

damage. The results of Shallice's study indicated that patients with anterior lesions made significantly more errors in completing the task than a group consisting of patients with posterior brain lesions and a group of normal controls. Shallice's control group comprised only 20 participants, and as a consequence offered only a limited estimation of the variance of the test in the normal population. Face validity implies that the Tower of London measures planning/problem-solving, since the subject has to perform the task in a limited number of moves, and hence has to work out how to solve the problem before attempting the solution. In addition, the task has been used in a number of clinical studies that have assumed that it is a measure of planning (Morris et al., 1990; Owen et al., 1990; Levin et al., 1991). However, the test has not been extensively investigated psychometrically in either clinical or normal populations.

The Six Element Test was designed by Shallice and Burgess (1991). They first used this test with 3 patients who had sustained traumatic brain injuries involving prefrontal structures. These data were assessed in a qualitative manner by Shallice and Burgess (1991), who suggested that the patients had more trouble with the task than 10 normal control participants. The Twenty Questions Test (Mosher & Hornsby, 1966) measures problem-solving and planning behavior in situations similar to those encountered in everyday life. The task follows the pattern of the well known party game, and has been adapted for experimental use with normal young children and adults, as well as detoxified long-term alcoholics and survivors of severe closed head injury (Mosher & Hornsby, 1966; Denny, 1973; Denny & Connors, 1974; Laine & Butters, 1982; Klouda & Cooper, 1990).

The Complex Figure Test of Rey (1941) was originally designed to evaluate memory but because of its complexity it has proved useful in evaluating the skills of planning, organizing, and assembling complex visual information (Goodglass & Kaplan, 1979; Binder, 1982; Klicpera, 1983; Waber & Holmes, 1986; Heinrichs & Bury, 1991).

METHOD

Research Participants

A sample of 130 neurologically normal individuals ages 17 to 55 years participated in this experiment. They were recruited from the student body of the University of Newcastle, New South Wales, Australia and from the local population of Newcastle. The sample comprised 56 men (M age = 25 years, SD = 7 years) and 74 women (M age = 26 years, 9 months; SD = 9 years), including 12 left-handers and 118 right-handers. The age group recruited, that of young adults, was selected in order to be comparable to head-injured populations, which show an age profile with a peak incidence in the young adult age group (Tate, 1989). All participants were screened during a comprehensive history taking to exclude people with a history of head injury and neurological disorder of the central nervous system.

Materials and Procedures

Eight neuropsychological tests of memory, attention, sequencing and planning/problem-solving were administered during a 2- to 3-hr test session. This report focuses on data from the following tasks: (1) the Tower of London Test (Shallice, 1982), (2) the Six Element Test (Shallice & Burgess, 1991); (3) the Twenty Questions Test (Mosher & Hornsby, 1966), and (4) the Complex Figure Test (Rey, 1941).

The Tower of London Test (Shallice, 1982) requires a subject to rearrange colored beads on upright rods in order to achieve a given target configuration in a limited number of moves (Morris et al., 1990). Test material for the Tower of London Test consists of three rods of differing heights connected to a base. Three colored beads (red, green, and blue) can be threaded onto the rods. In addition to the original three-rod design, a modified version in which there are four beads (red, green, blue, and yellow) and four rods was also used. The rules of the task were the same as those described by Shallice (1982); however participants were given unlimited time to solve each problem. Two subsets of eight problems were devised, each containing 4 levels of difficulty (3-, 4-, 5-, and 6-move levels of difficulty) with two trials for each level of difficulty for each apparatus type (three or four rods). On the first set of trials for each apparatus type, subjects were instructed merely to solve the problems in the minimum number of moves. On the second set of trials, subjects were first required to inform the experimenter of the minimum number of moves they needed to solve the problem, and then proceed to demonstrate the solution. This created a *no-cue* and a *cue* condition, respectively. It was assumed that requiring subjects to plan out their solution before making the first move would increase their accuracy by reducing trial-and-error moves, a result found by Alhum-Heath and di-Vesta (1986), using the Tower of Hanoi.

The Tower of London also provides measures of the rate of information processing by using a procedure devised by Van Zomeren (1981). Each problem was divided into two components: planning (or decision) time, and solution (or movement) time. Planning time is measured in seconds from the end of initial viewing of the problem until the first bead clears a rod. Solution time is measured from when the first bead clears a rod until completion of the response as confirmed by the subject.

The following measures were collated from each condition and trial of the Tower of London Test: (1) the mean number of moves to solution, (2) the mean number of moves above the minimum, (3) the number of rule breaks, (4) planning time, and (5) movement time. For the cued condition an additional measure was included: the discrepancy between the reported and the actual number of moves required.

The Six Element Test was designed by Shallice and Burgess (1991) to reflect the type of planning and decision-making inherent in everyday life. The test requires subjects to organize their behavior over time in order to carry out six open-ended tasks within 15 min. The six tasks are divided

into two sets of three: (1) dictating routes of journeys, (2) carrying out arithmetic problems, and (3) writing down the names of approximately 100 pictures of objects. To complete the task successfully, participants need to divide their time evenly between the tasks, to complete as much as possible of every task, and to adhere to a rule that states that they should not do two tasks of the same type one after the other. Test variables for this task consist of: (1) the number of subtasks attempted, (2) the maximum time on any subtask, (3) the number of task changes, (4) the number of rule breaks, and (5) an overall score measuring the number of task changes divided by the number of subtasks tackled.

The Twenty Questions Test was designed by Mosher & Hornsby (1966) to measure problem solving and planning behavior in situations similar to those encountered in everyday life. For this task subjects are required to ask questions of the experimenter in order to determine what subject, in this case an animal; the experimenter is thinking of. The experimenter can only respond "yes" or "no," and the object of the test is to guess the animal with the fewest number of questions. The questions are classified as follows (see Klouda & Cooper, 1990): as (1) *constraint-seeking (CS)*: general questions designed to eliminate many possibilities (e.g., "Is it a mammal?"); (2) *hypothesis scanning (HS)*: specific questions or guesses (e.g., "Is it a cat?"); and (3) *pseudo-constraint-seeking (PC)*: questions phrased like a constraint-seeking question but actually referring to only one animal (e.g., "Does it bark?"). Each participant was given two versions of the game, and the solutions were *chicken* and *elephant*. If the participant failed to solve the task within 20 questions, the game continued until it was either solved correctly, or the participant requested that it be discontinued. For scoring purposes, the number and type of question were scored for the first five questions only, since some subjects guessed correctly within five questions (Klouda & Cooper, 1990). In addition an overall score was calculated, taking into account the number of questions asked in both trials, and whether or not the participant solved the problem.

The Complex Figure Test (Rey, 1941) requires test-takers to copy the complex figure, and then without warning reproduce it from memory at a later stage. Although originally devised as a measure of memory, it has also proved useful in evaluating the skills of planning and organization. Shorr et al., (1992) developed a scoring system to quantify the use of organized strategy for copying the numerous subelements and isomorphic features from the same perceptual category. For this scoring procedure the figure is divided into subunits, and each subunit is divided into junctions where breaks in drawing can occur. The scoring system used in the present study was the same as that described by Shorr et al. (1992). Shorr et al. tested their scoring procedure on a mixed group of 50 neuropsychiatric patients, and found that the perceptual cluster score in the copy condition was a better indicator of memory performance than the traditional Taylor copy score (Taylor, 1959). However, Shorr et al. (1992) did not use a normal control group in their study, and hence there is no indication of the meaning of the score for a normal population. Only the copy score was used in this study, since the interest was in planning rather than memory.

RESULTS

Inspection of the frequency distributions of the variables showed markedly nonnormal distributions for all four tests. Means, standard deviations and ranges are reported in Table 1 and Table 2. Therefore the variables were normalized by transform procedures using the (SPSS Inc., 1983) *normal* command.

Initially, one-way analyses of variance were performed on all test variables in order to identify any effect of gender. For Game 1 of the Twenty Questions Test a main effect for gender emerged for the number of constraint-seeking questions asked [$F(1,126) = 5.87$, $p < .05$]. Men (Transformed $M = 1.67$, Raw $M = 4.6964$) were found to have asked more constraint-seeking questions than women (Transformed $M = -.63$, Raw $M = 4.64$). This result was not replicated in Game 2 where both men (Transformed $M = -.2750$, Raw $M = 4.79$) and women (Transformed $M = -.0127$, Raw $M = 4.71$) asked fewer constraint-seeking questions. On the Complex Figure Test, it was found that female participants (Transformed $M = 1.17$, Raw $M = 17.62$) scored significantly higher than men (Transformed $M = -5.01$, Raw $M = 17.02$) on the Shorr cluster score [$F(1,126) = 4.50$, $p < .05$], which suggests that the women used a better organized strategy than men when copying the figure.

Repeated measures analyses of variance ($2 \times 2 \times 4 \times 2$) were conducted on the Tower of London Test data in order

Table 1. Means, standard deviations, and range for variables from the Six Element Test, the Twenty Questions Test, and the Complex Figure Test

Variable	*M*	*SD*	Range
Six Element			
No. Subtasks	5.34	1.01	2–6
Max Time (s)	272.4	125.8	84–826
No. Changes	6.13	3.02	1–23
Rule Breaks	0.27	0.72	0–4
Overall score	1.11	0.46	0.5–3.8
20 Questions			
Mean No. Asked	14.52	8.41	5–56
Mean % CS	94.15	12.52	0–100
Mean % HS	4.62	10.27	0–60
Mean % PC	1.23	7.92	0–100
Total score	1.62	0.60	0–2
Complex Figure Test			
Copy time (s)	135.2	52.7	43–331
Taylor score	35.6	0.81	32–36
Shorr ratio	0.87	0.15	0.25–1

CS = constraint-seeking questions, HS = hypothesis scanning questions, PC = pseudo-constraint-seeking questions.

https://doi.org/10.1017/S1355617797001082 Published online by Cambridge University Press

Table 2. Means, standard deviations, and range for Tower of London variables

Level	Mean no. of moves	Mean above minimum	Planning time (s)	Movement time (s)	Rule breaks
3-Rod apparatus					
3 Moves					
M	3.19	0.19	5.40	5.53	0.01
SD	0.40	0.40	2.21	1.94	0.07
Range	3–6	0–3	2–13	3–13	0–0.5
4 Moves					
M	5.79	1.79	9.52	14.52	0.01
SD	1.88	1.88	5.18	9.80	0.07
Range	4–14	0–10	2–30	4–52	0–0.5
5 Moves					
M	6.95	1.95	15.80	19.28	0.01
SD	1.59	1.60	11.56	14.12	0.08
Range	5–14	0–9	4–65	7–100	0–0.75
6 Moves					
M	8.24	2.25	26.97	23.11	0.03
SD	1.69	1.71	18.05	11.27	0.20
Range	6–14	0–8.5	4–79	8–62	0–2
4-Rod apparatus					
3 Moves					
M	3.11	0.11	4.44	5.0	0.0
SD	0.27	0.27	1.70	1.43	0.2
Range	3–4.5	0–1.5	2–12	3–14	0–0.25
4 Moves					
M	4.18	0.18	6.04	6.67	0.0
SD	0.37	0.37	3.27	1.52	0.02
Range	4–6	0–2	2–34	4–14	0–0.25
5 Moves					
M	5.77	0.77	13.76	11.82	0.0
SD	0.66	0.66	8.46	3.99	0.02
Range	5–8	0–3	3–65	6–34	0–0.25
6 Moves					
M	6.37	0.37	17.57	12.83	0.01
SD	0.39	0.39	11.24	4.01	0.04
Range	6–7.5	0–1.5	4–70	7–32	0–0.25

to examine the effects of gender, cue condition, levels of problem difficulty, and apparatus type (three or four rods) on the performance of this test. A main effect for difficulty was found for planning time [$F(3,378) = 3.39$, $p < .05$]. *Post hoc* analyses failed to reveal any significant differences among the means; however, a trend was observed in which participants took greater planning time (Transformed $M = .573$, Raw $M = 89.1$) for problems requiring a minimum of six moves to solution than problems requiring five-move solutions (Transformed $M = -4.79$, Raw $M = 59.1$). No main effect for cue condition was found, which indicated that asking participants to plan their solution prior to making the first move did not influence performance. For the variable measuring the discrepancy between the reported number of moves to solve and the actual number of moves needed for solution in the cue condition, a three-way interaction between sex, difficulty and type of apparatus was found [$F(3,378) = 2.94$, $p < .05$]. *Post hoc* analyses revealed that when using the three-rod apparatus, women (Transformed $M = .90$, Raw $M = 2.60$) had a significantly greater discrepancy than men (Transformed $M = -.33$, Raw $M = 1.83$) on problems of six-move difficulty. It was also found that on the three-rod apparatus with problems of six-move difficulty, women (Transformed $M = .90$, Raw $M = 2.60$) demonstrated a greater discrepancy than men for both the four-rod apparatus with problems of three-move difficulty (Transformed $M = -.06$, Raw $M = .13$) and for the three-rod apparatus with problems of four-move difficulty (Transformed $M = .20$, Raw $M = .63$). However, the greater discrepancy shown by women in these interactions was not shown overall, since a direct comparison of male and female discrepancy scores was not significant [$F(1,126) = 1.33$, $p = .251$].

The structural equation modeling procedures reported below employed the LISREL 8 routines of Jöreskog and Sörbom (Jöreskog & Sörbom, 1993). All LISREL models were

https://doi.org/10.1017/S1355617797001082 Published online by Cambridge University Press

estimated by analyzing a matrix of product–moment correlations using Maximum Likelihood estimation (MLE). The maximum likelihood estimates are obtained by means of an iterative procedure that minimizes a particular fit function by successively improving the parameter estimates (Jöreskog & Sörbom, 1993). The sample size here was too small to use the generally weighted least squares estimation (WLS) with an asymptotic correlation matrix to control for the effects of skewed and kurtotic distributions. It must be noted that possible violations of the assumptions of normality may occur despite efforts to create more normal distributions by transforming the data to normalized variables.

The data were analyzed in two steps. First, variables were selected from each test to reflect the most intuitively appropriate measures of the latent variable Planning/Problem-Solving. One variable was chosen from each test for the Six Element Test, the Twenty Questions Test, and the Complex Figure Test. The measures were the Six Element overall score, the Twenty Questions overall score, and for the Complex Figure task, the Shorr et al. ratio score. Summary statistics for these variables are shown in Table 3.

From the 88 Tower of London variables, 8 were chosen as the ones that intuitively seemed most likely to be sensitive to planning/problem-solving. These were planning time and mean number of moves above the minimum for the three-rod and four-rod apparatus across both cue and no-cue conditions at the six-move level of difficulty (Table 3). A one-factor congeneric model of these variables was tested to determine whether these Tower of London variables reflected one underlying construct, the latent variable Planning/Problem-Solving. This model failed to meet the admissibility criteria, indicating that some of the parameter estimates were unacceptable. This failure to confirm the admissibility of the designated measures was unexpected, since these variables were chosen as ones that would be the most sensitive to the construct of planning/problem-solving. In view of this unexpected finding with confirmatory approaches, it was decided to perform an exploratory factor analysis on the results in order to see if any pattern of parameter estimates could be identified. A principal components analysis with oblique rotation was performed on the eight selected Tower of London variables. Four factors with eigenvalues greater than 1 were extracted, accounting for 59.7% of the variance (Table 4).

The first factor, accounting for 18% of the variance, had primary factor loadings from two variables. Both were from the four-rod apparatus in the no-cue condition; one measuring planning time, and the other, mean number of moves above minimum. The second factor accounted for 15.3% of the variance and had primary loadings from two variables, both for the cued condition, but one for the three-rod apparatus and the other for the four-rod apparatus. The last two factors accounted for 13.9% and 12.5% of the variance, respectively. The third factor had loadings from three variables: two mean number of moves above minimum, and one planning time variable from different cue conditions and apparatus types. The final factor showed loadings from one variable: mean number of moves above minimum for the three-rod apparatus at the six-move level of difficulty in the cue condition. The results of the principal components analysis revealed no consistent or easily interpretable pattern of loadings. There was a scattering of planning time and mean number of moves above minimum variables for different cue conditions and apparatus types across factors. Since no consistent pattern of factor loadings could be identified from the eight chosen variables, it was decided to explore all the other Tower of London variables in an attempt to identify other variables that might provide a more reliable measure of planning. Hence, a second principal components analysis was conducted on all 88 Tower of London test variables. Thirty-five factors with eigenval-

Table 3. Means and standard deviations for transformed variables in LISREL analysis

Variable	*M*	*SD*
6 Element overall	.032	1.33
20 Questions overall	−1.31	21.14
RCF Shorr ratio	.101	.86
Tower of London		
Planning time		
No-cue/3-rod	−.36	17.1
Cue/3-rod	−.53	41.2
No-cue/4-rod	−.47	7.5
Cue/4-rod	6.55	33.3
Mean moves above minimum		
No-cue/3-rod	2.64	2.02
Cue/3-rod	.13	3.6
No-cue/4-rod	.08	1.1
Cue/4-rod	.02	1.1

Table 4. Factor analysis for eight selected Tower of London variables

Variable	Factor 1	Factor 2	Factor 3	Factor 4
X1	−.724			
X2	.704			
X3		−.782		
X4		.710		
X5			.781	
X6			−.576	
X7			.470	
X8				.925
Eigenvalue	1.44	1.22	1.11	1.00
% of variance	18.0	15.3	13.9	12.5

All Tower of London variables are for the 6-move level of difficulty and are identified as follows: Planning time: X1 = 4-rod, no-cue; X3 = 4-rod, cue; X4 = 3-rod, cue; X6 = 3-rod, no-cue. Mean moves above minimum: X2 = 4-rod, no-cue; X5 = 4-rod, cue; X7 = 3-rod, no-cue; X8 = 3-rod, cue.

https://doi.org/10.1017/S1355617797001082 Published online by Cambridge University Press

ues greater than 1 were extracted. The first factor had an eigenvalue of 3.55 and accounted for 4% of the variance.

The variables loading onto Factor 1 showed no consistent pattern of influence with variables measuring different scores, different cue conditions and different apparatus types. The conclusion from these Tower of London Test results is that these data show such a large degree of variance that no underlying structure of responses emerges. Certainly, the partial correlations do not reveal the simple latent variable structure of a single construct, Planning/Problem-Solving, which had been hypothesized. In addition, it must be noted that a 35-factor solution involving 130 participants does not meet the minimal standards for an acceptable subject-to-variable ratio (Francis, 1988).

The second step in the structural equation modeling involved an investigation of the relationships between the test variables from the four tests of Planning/Problem-Solving and the underlying structure of latent variables using the LISREL measurement model. In order to include the Tower of London measures in light of the PCA results, it was decided to revert to the initially intuitively selected measures; hence all the eight Tower of London variables initially selected were included. The reason for including these variables was that they possibly could show relationships with measures from the other tests even though they showed no consistent pattern of relationships themselves.

The first hypothesis to be tested using the LISREL 8 program was that Planning is a unidimensional construct underlying the four tests and their associated variables (Figure 1). In Figure 1 the Tower of London variables are represented as Variables x1 to x8, while the variables from the CFT, the Twenty Questions Test, and the Six Element Test are Variables x9, x10, and x11, respectively.

The initial model did not fit the data well. Although the χ^2/df ratio was less than 2.00, the other indices of admissibility were not within the acceptable limits.

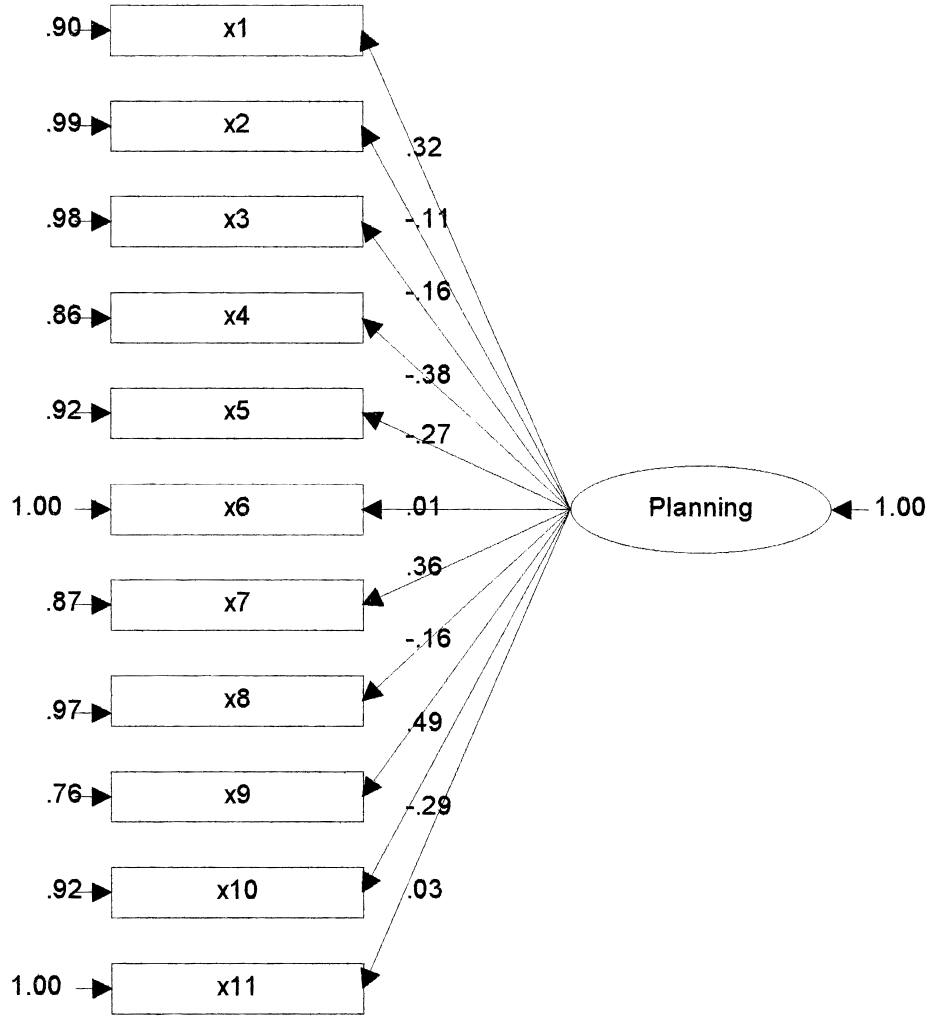

Fig. 1. Initial one-factor model of Planning/Problem-Solving. Test variables are as follows: Tower of London, planning time, X2 = 3-rod, no-cue; X4 = 3-rod, cue; X6 = 3-rod, no-cue; X8 = 4-rod, cue; mean no. moves above minimum, X1 = 3-rod, no-cue; X3 = 3-rod, cue; X5 = 4-rod, no-cue; X7 = 4-rod, cue; X9 = Complex Figure Test Shorr ratio score; X10 = Twenty Questions overall score; X11 = Six Element overall score.

https://doi.org/10.1017/S1355617797001082 Published online by Cambridge University Press

If the goodness-of-fit of a model is inadequate, then various indices can be used to help detect the source of misfit. These include squared multiple correlations (R^2), T values, standard errors, and modification indices (MI; Byrne, 1989). In general reestimation of a model should not be attempted, since the objective is to confirm a set of hypothesized relationships, rather than to explore a set of possible factor loadings. However for this initial model many of the R^2 values for the Tower of London variables were very low, and ranged from .01 to .24. This suggested that some of the indicators used in the model were not accurately measuring Planning/Problem-Solving. In addition, the previously established unreliability of the Tower of London variables also indicated that model reestimation would be appropriate in this case. Thus the initial model was reestimated four times, each time removing Tower of London variables with unacceptably low R^2 values.

This final one-factor model contained seven observed variables; four from the Tower of London (x1 to x4) and one from each of the remaining three tests (x5, x6, and x7). This model is shown in Figure 2.

All goodness-of-fit measures except the Bentler and Bonnett Normed Index indicated that this model was an acceptable fit to the observed data (Bentler & Bonnett, 1980). However, examination of the R^2 values revealed low scores ranging from .01 to .36. Furthermore, the two overall measures from the Six Element Test and the Twenty Questions Test did not have significant T values. It was therefore concluded that these models did not adequately represent the observed data.

The second hypothesis was that Planning/Problem-Solving is not a single construct but, a two-factor construct, with the Tower of London measuring a different aspect of planning from the other three tests (Figure 3). In Figure 3 the Tower of London variables are represented as x1 to x8, and the CFT, the Twenty Questions Test, and the Six Element Test are represented as x9, x10 and x11, respectively.

This model failed to adhere to all of the goodness-of-fit indices. Examination of the T values for the variables revealed that only one of the variables was significant in the model: the overall score for the Twenty Questions Test (x10). The standard errors for some of the Tower of London variables were unacceptably large, although the standard errors associated with the other test variables were acceptable.

Based on these results no further attempts were made to reestimate the model. Jöreskog and Sörbom (1993) suggest that any model that gives unreasonable results should be eliminated from further consideration.

The final hypothesis tested was that Planning/Problem-Solving is a four-factor construct with each test measuring a different aspect of the construct and loading on separate underlying latent variables (Figure 4). In Figure 4 the Tower of London variables are labelled as x1 to x8, and the variables from the CFT, the Twenty Questions Test and the Six Element Test are labelled as x9, x10, and x11, respectively.

This model also failed to meet the goodness-of-fit criteria. Furthermore examination of T values revealed that none of the variables were significant variables to the model. However, the standard errors for the variables were acceptable.

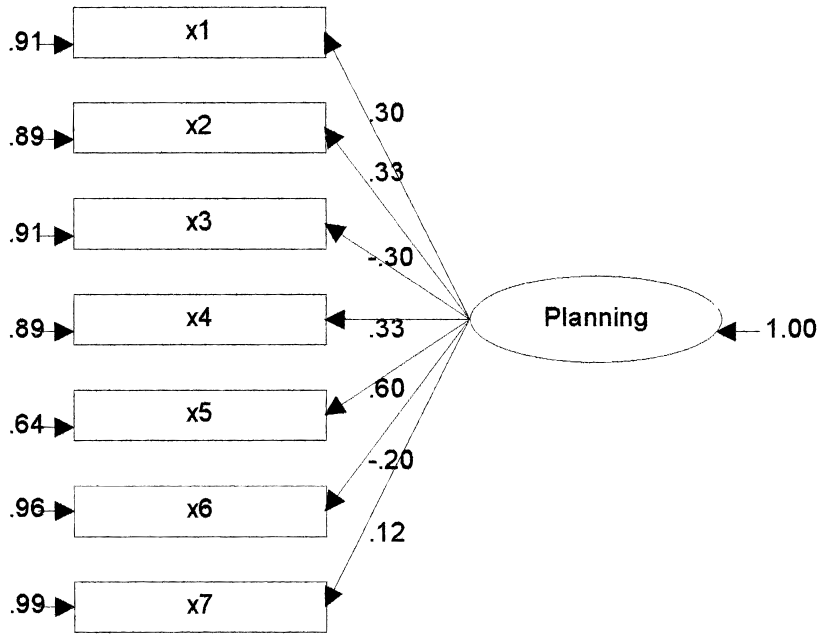

Fig. 2. Final one-factor model of Planning/Problem-Solving. Test variables are as follows: Tower of London, planning time, X2 = 3-rod, cue; mean no. moves above minimum, X1 = 3-rod, no-cue; X3 = 4-rod, no-cue; X4 = 4-rod, cue; X5 = Complex Figure Test Shorr ratio score; X6 = Twenty Questions overall score; X7 = Six Element overall score.

https://doi.org/10.1017/S1355617797001082 Published online by Cambridge University Press

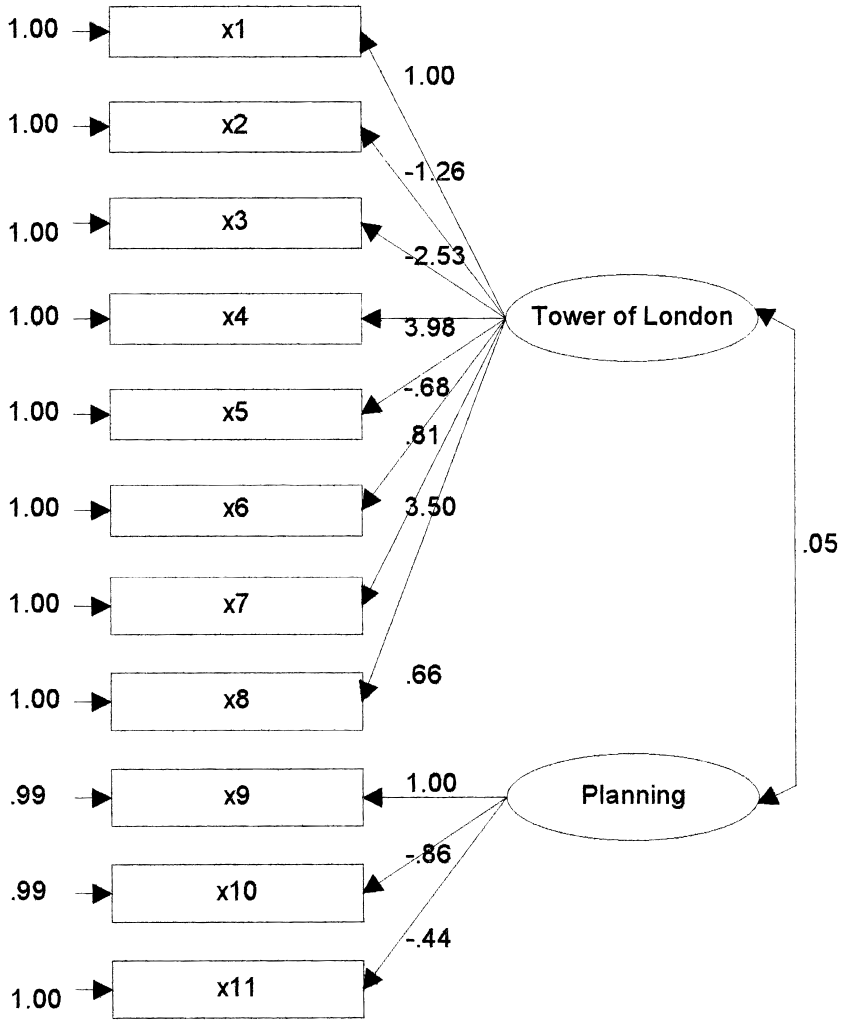

Fig. 3. Two-factor model of Planning/Problem-Solving. Test variables are as follows: Tower of London, planning time, X2 = 3-rod, no-cue; X4 = 3-rod, cue; X6 = 3-rod, no-cue; X8 = 4-rod, cue; mean moves above minimum, X1 = 3-rod, no-cue; X3 = 3-rod, cue; X5 = 4-rod, no-cue; X7 = 4-rod, cue; X9 = Complex Figure Test Shorr ratio score; X10 = Twenty Questions overall score; X11 = Six Element overall score.

The correlations between the four latent constructs were also not significant (Table 5). This implies that the four tests are measuring different, unrelated constructs. Moreover, the overall poor fit of the model further implies that the relationships are more complicated and structured differently from the pattern of relationships that was hypothesized.

None of the models proposed to explain the structure of the construct Planning/Problem-Solving in relation to the four tests fitted the observed data in an acceptable way.

Table 5. *T* values and standard errors (SE) for associations between latent variables

Latent variable	Tower	CFT	20 Q's	6 Element
Tower	—			
CFT	.77 (.06)	—		
20 Q's	−.77 (.06)	−.86 (.09)	—	
6 Element	−.65 (.02)	1.93 (.09)	1.03 (.09)	—

Tower = Tower of London Test, CFT = Complex Figure Test, 20 Q's = Twenty Questions Test, 6 Element = Six Element Test.

DISCUSSION

This study examined the psychometric properties and relationships among the Tower of London Test, the Twenty Questions Test, the Six Element Test and the Complex Figure Test. The results of the present study raise a number of important issues that require comment. In particular, the difficulty of measuring the construct planning/problem-solving and the problems of the psychometric structure of the Tower of London Test need to be addressed.

The LISREL structural equation modeling approach has been used widely in the area of personality research, and in

https://doi.org/10.1017/S1355617797001082 Published online by Cambridge University Press

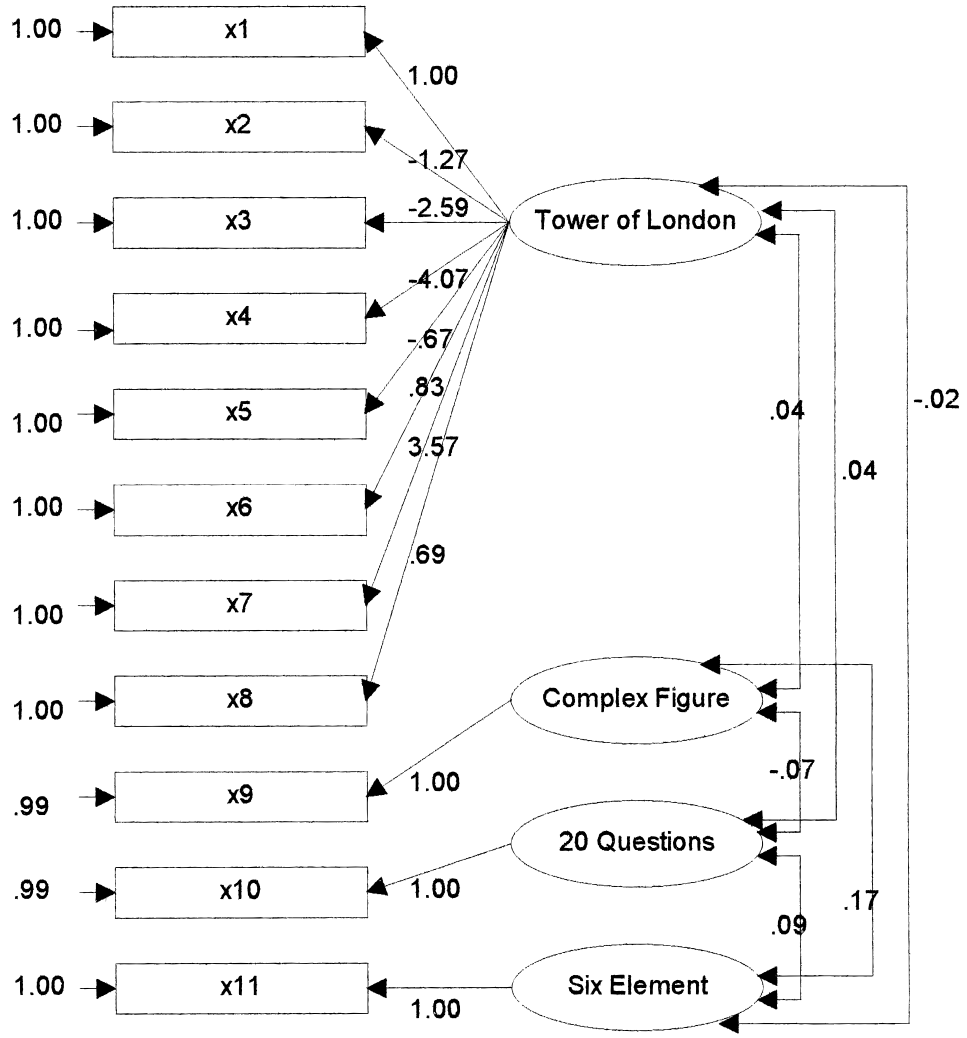

Fig. 4. Four-factor model of Planning/Problem-Solving. Test variables are as follows: Tower of London, planning time, X2 = 3-rod, no-cue; X4 = 3-rod, cue; X6 = 3-rod, no-cue; X8 = 4-rod, cue; mean moves above minimum, X1 = 3-rod, no-cue; X3 = 3-rod, cue; X5 = 4-rod, no-cue; X7 = 4-rod, cue; X9 = Complex Figure Test Shorr ratio score; X10 = Twenty Questions overall score; X11 = Six Element overall score.

the analysis of questionnaire data. Although the techniques of structural equation modeling have been known for several years, they have not been used to investigate neuropsychological concepts.

The results of this study revealed that an adequate model for the construct Planning/Problem-Solving could not be estimated. None of the hypothesized models were consistent with the observed data. This raises issues concerning the validity of the concept of planning/problem-solving. Four possible explanations for this result are proposed: (1) the tests chosen did not measure Planning/Problem-Solving, (2) the test measures chosen were not sensitive to planning, (3) the sample size in the study was too small to use a structural equation modeling approach appropriately, or (4) Planning/Problem-Solving as a construct is so complex that none of the tests accurately or validly measure it.

First, the face validity of these tests, which leads to their wide clinical use, suggests that they do measure some aspect of planning/problem-solving. In addition, research has shown that they are sensitive to dysfunction of frontal executive abilities in clinical populations, even though sample sizes in these studies usually have been small (Shallice, 1982; Owen et al., 1990; Shallice & Burgess, 1991). So perhaps it is the case that these tests are only sensitive to planning/problem-solving abilities in clinical populations, and not in neurologically normal populations. These tests simply may not be sensitive enough to detect variations in performances that are narrowly distributed in normal subjects. This could be particularly true for the Twenty Questions Test, where 94% of questions asked were constraint-seeking questions. However, the absence of ceiling effects for the Tower of London Test, particularly for the three-rod apparatus, implies that, even in normal subjects, planning and problem-solving capabilities are required. Furthermore, it could be proposed that the exclusion of participants over the age of 55 years from the study may have

https://doi.org/10.1017/S1355617797001082 Published online by Cambridge University Press

further restricted the variations in executive performance evident in the normal population. This is a possibility; however, the inclusion of older individuals also creates methodological problems such as age-related decline on visuospatial and memory abilities. Ideally, future research would benefit from examining tests of executive functions in a clinical population of patients with identified executive dysfunction. However there are no standardized quantitative means of identifying patients with executive dysfunction. In addition, not all patients with frontal lobe damage have poor executive functions, and this may be a factor contributing to the lack of conclusive findings concerning executive functions and frontal lobe damage.

Second, it could be proposed that the measures chosen from each test to represent planning/problem-solving were not the most sensitive to planning/problem-solving. The Tower of London measures of planning reported in the literature are not consistent. However, studies by Ponsford and Kinsella (1992) and Owen et al. (1990) have both found planning time and the mean number of moves to be sensitive to frontal brain damage. Previous research on the Six Element Test has been purely qualitative, and while differences were observed between head injured patients and normal controls, these differences did not indicate which of the measures were most sensitive to planning. As a consequence, the current study used a composite score, taking into account the number of task changes and the number of tasks tackled. This appeared intuitively to be the most likely measure to be sensitive to planning based on face validity. Similarly, for the Twenty Questions Test a composite overall score was used that took into account both the number of questions asked and the ability to solve each problem. Goldstein and Levin (1991) found that head injured patients required more questions than normal controls to guess an item on a pictorial version of the task. So again, the choice of measure seemed reasonable and defensible, based on face validity and reported use with clinical populations. For the Complex Figure Test, previous research has shown the Shorr et al. (1992) scoring system to be more sensitive than the traditional Taylor score. It could be argued that this score reflects visuospatial learning rather than planning/problem-solving, although this would seem unlikely on the basis of the results of Shorr et al. (1992). The scoring technique developed by Shorr et al. (1992) measures the way a person constructs the figure in terms of perceptual subelements. In contrast, the Taylor score only measures the presence or absence of various lines. It is more likely that the Taylor score is indicative of visuospatial learning, whereas the Shorr et al. score is more indicative of some form of planning or organization. Nevertheless, the measures were chosen on the basis of previous research and face validity, and yet none of the models provided a good fit to these data.

Third, regarding sample size, five LISREL 8 indices were used in this study to assess the acceptability of estimated models. Of these five indices, two were independent of sample size: the Goodness-of-Fit Index (GFI), and the Adjusted Goodness-of-Fit Index (AGFI; Byrne, 1989; Jöreskog & Sörbom, 1993). The GFI indicates the relative amount of variance and covariance jointly explained by the model. The AGFI is a similar measure, however it takes into account the degrees of freedom in the model. Both indices range from 0 to 1, with a value closer to 1 representing a good fit (Byrne, 1989). An AGFI of greater than or equal to .94 was used in this study to indicate an acceptable fit between the model and the data. This value is recommended by Byrne (1989); however, no single value is standardly reported in the literature. In addition, the LISREL results concerning the Tower of London were replicated through factor analyses. It must be admitted that a smaller-than-ideal sample size may have affected the results, but the relatively conservative use of the fit indices effectively counters this problem.

Finally, it could be suggested that planning/problem-solving is such a complex construct that it is not easily measured by any one test. If this is the case then planning/problem-solving might be better measured by breaking it down into subelements such as motor planning, logical sequencing and goal-orienting behavior. If these subelements all measure an aspect of the same higher-order construct, one would expect them to be correlated. In the current study not only was it found that the observed data could not be adequately explained with a four-factor model (that is, one factor for each test), but in addition, the four factors were not significantly correlated. These findings imply that the relationships between the test measures and their underlying construct or constructs is far more complicated and structured differently from the patterns of relationships that had been hypothesized. If this is the case, then serious reconsideration must be given to the continued use of these tests as estimates of planning and problem-solving deficits in clinical populations.

Although the Tower of London Test is widely used as a clinical measure of frontal executive deficits, it was found here to neither reliably nor accurately measure planning/problem-solving. If the Tower of London test fails to measure planning/problem-solving then what does it measure? Thirty-five factors were extracted in a factor analysis with no consistent or easily interpretable pattern of results. If this test is not validly and accurately measuring anything meaningful within a normal population, it begs the question as to what clinicians are measuring when they administer the test to a brain-injured client. It could be argued that the test is appropriate for a brain injured population but not a normal population. Such an argument might be possible if a ceiling effect were observed in the normal population. However, the test does allow for different levels of difficulty, and the measures chosen for analysis here show levels of variability that deny the possibility of a ceiling effect.

These results found in our normal adult population contrast with the findings of Levin et al. (1991) in a sample of normal children ages 7 to 15 years. They found the Tower of London sensitive to developmental changes in children. In particular, children between 13 and 15 years of age solved more problems on the first trial and required fewer trials to solve the test than children ages 7 and 8 years. Moreover, in

a principal components analysis with variables from the Tower of London, Wisconsin Card Sorting Test (Grant & Berg, 1948), verbal fluency (Benton & Hamsher, 1976), design fluency (Jones-Gotman & Milner, 1977), California Verbal Learning Test–Children's Version (CVLT; Delis et al., 1986), the Twenty Questions Test (Denny & Denny, 1973), and the Go-No Go task (Drewe, 1975), the Tower of London was found to load on a separate factor. The Twenty Questions Test was found to load on another factor with variables from the CVLT and verbal fluency. Again then, the Tower of London Test seems not to correlate highly with other so-called planning tests, and the cross-sectional design and small sample sizes make generalization of their results difficult, as the authors acknowledge. Obviously more research is required on the Tower of London Test in both normal samples and brain-injured subjects in order to resolve its apparent lack of validity.

While previous research has assumed that various neuropsychological tests accurately and validly measure executive functions, the results of our study imply that this is not the case. Indeed we would argue that until further work is carried out on the design and norming of such tests, clinicians should be circumspect in their use and interpretation of the Tower of London Test and other tests purported to measure executive functions.

ACKNOWLEDGMENTS

Funding for this research was supported in part by an Australian Post-Graduate Research Award to Ms. Kristine Kafer, and in part by the Department of Psychology, University of Newcastle. The authors gratefully acknowledge the assistance of Anthony Ruge in the data collection, Dr. Philip Holmes-Smith in the LISREL 8 analyses; and Drs. Norman Kafer and Stephen Provost for their comments on drafts of this manuscript. Portions of this paper were presented at the Australian Psychological Society Clinical Neuropsychologists 10th National Conference, Sydney, NSW, Australia, September 1994.

REFERENCES

Alhum-Heath, M.E. & di-Vesta, F.J. (1986). The effect of conscious controlled verbalization of a cognitive strategy on transfer in problem solving. *Memory and Cognition, 14*, 281–285.

Bentler, P.M. & Bonnett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606.

Benton, A.L. & Hamsher, K. (1976). *Multilingual aphasia examination*. Iowa City, IA: University of Iowa Press.

Binder, L.M. (1982). Constructional strategies on complex figure drawings after unilateral brain damage. *Journal of Clinical Neuropsychology, 4*, 51–58.

Byrne, B.M. (1989). *A primer of LISREL.* New York: Springer Verlag.

Delis, D.C., Kramer, J.H., Kaplan, E., & Ober, B.A. (1986). *The California Verbal Learning Test* (Research ed.). New York: The Psychological Corporation.

Denny, D.R. (1973). Reflection and impulsivity as determinants of conceptual strategy. *Child Development, 44*, 614–623.

Denny, N.W. & Connors, G.J. (1974). Altering the questioning strategies of preschool children. *Child Development, 45*, 1108–1112.

Denny, D.R. & Denny, N.W. (1973). The use of classification for problem-solving: A comparison of middle and old age. *Developmental Psychology, 9*, 275–278.

Drewe, E.A. (1975). Go-no go learning after frontal lobe lesions in humans. *Cortex, 11*, 8–16.

Francis, D.J. (1988). An introduction to structural equation models. *Journal of Clinical and Experimental Neuropsychology, 10*, 623–639.

Goldstein, F.C. & Levin, H.S. (1991). Question-asking strategies after severe closed head injury. *Brain and Cognition, 17*, 23–30.

Goodglass, H. & Kaplan, E. (1979). Assessment of cognitive deficit in the brain-injured patient. In M.S. Gazzaniga (Ed.), *Handbook of behavioral neurobiology: Vol. 2. Neuropsychology* (pp. 3–22). New York: Plenum Press.

Grant, D.A. & Berg, E.A. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new response in a Weigl type card sorting problem. *Journal of Experimental Psychology, 38*, 404–411.

Heinrichs, R.W. & Bury, A. (1991). Copying strategies and memory on the Complex Figure Test in psychiatric patients. *Psychological Reports, 69*, 223–226.

Jones-Gotman, M. & Milner, B. (1977). Design fluency: The inventions of nonsense drawings after focal cortical lesions. *Neuropsychologia, 15*, 653–674.

Jöreskog, K.G. & Sörbom, D. (1993). *New features in LISREL 8.* Chicago: Scientific Software.

Klicpera, C. (1983). Poor planning as a characteristic of problem-solving behavior in dyslexic children. *Acta Paedopsychiatrica, 49*, 73–82.

Klouda, G.V. & Cooper, W.E. (1990). Information search following damage to the frontal lobes. *Psychological Reports*, *67*, 411–416.

Laine, M. & Butters, N. (1982). A preliminary study of the problem-solving strategies of detoxified long-term alcoholics. *Drug and Alcohol Dependence, 10*, 235–242.

Levin, H.S., Culhane, K.A., Hartmann, J., Evankovich, K., Mattson, A.J., Harward, H., Ringholz, G., Ewing-Cobbs, L., & Fletcher, J.M. (1991). Developmental changes in performance on tests of purported frontal lobe functioning. *Developmental Neuropsychology, 7*, 377–395.

Lezak, M.D. (1983). *Neuropsychological assessment* (2nd ed.). New York: Oxford University Press.

Luria, A.R. (1973). *The working brain: An introduction to neuropsychology.* Harmondsworth, U.K.: Penguin.

Morris, R.G., Downes, J.J., & Robbins, T.W. (1990). The nature of the dysexecutive syndromes in Parkinson's disease. In K.J. Gilhooly, M.T.G. Keane, R.H. Logie, & G. Erdos (Eds.), *Lines of thinking* (Vol. 2., pp. 247–258). West Sussex, UK: Wiley.

Mosher, F.A. & Hornsby, J.R. (1966). On asking questions. In J.S. Bruner, R.R. Olver, P.M. Greenfield, J.R. Hornsby, H.J. Kenney, M. Maccoby, N. Modiano, F.A. Mosher, D.R. Olson, M.C. Potter, L.C. Reich, & A. Sonstroem (Eds.), *Studies in cognitive growth* (pp. 86–102). New York: Wiley.

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*, 1021–1034.

Ponsford, J. & Kinsella, G. (1992). Attentional deficits following closed-head injury. *Journal of Clinical and Experimental Neuropsychology, 14*, 822–838.

Pusakulich, R.L. (1992). Using a model of cognitive function to plan cognitive treatment. In C.J. Long & L.K. Ross (Eds.), *Handbook of head trauma: Acute care to recovery. Critical issues in neuropsychology* (pp. 91–105). New York: Plenum Press.

Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatiques [The psychological examination in cases of traumatic encephalopathy]. *Archives de Psychologie, 28*, 286–340.

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

Shallice, T. & Burgess, P.W. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain, 114*, 727–741.

Shorr, J.S., Delis, D.C., & Massman, P.J. (1992). Memory for the Rey–Osterrieth figure: Perceptual clustering, encoding and storage. *Neuropsychology, 6*, 43–50.

Sohlberg, M.M. & Mateer, C.A. (1989). *Introduction to cognitive rehabilitation: Theory and practice.* New York: Guilford Press.

SPSS Inc. (1983). *SPSS User's Guide*. New York: McGraw-Hall.

Tate, R.L. (1989). *Differential Patterns of Neuropsychological Impairment in Brain Injured Patients.* Unpublished doctoral dissertation, University of Newcastle, NSW, Australia.

Taylor, E.M. (1959). *Psychological appraisal of children with cerebral deficits.* Cambridge, MA: Harvard University Press.

Van Zomeren, A.H. (1981). *Reaction time and attention after closed head injury.* Lisse, The Netherlands: Swets & Zeitlinger.

Vilkki, J. & Holst, P. (1991). Mental programming after frontal lobe lesions: Results on digit symbol performance with self-selected goals. *Cortex, 27*, 203–211.

Waber, D.P. & Holmes, J.M. (1986). Assessing children's memory productions of the Rey–Osterrieth complex figure. *Journal of Clinical and Experimental Neuropsychology, 8*, 563–580.

Wong, B.Y.L. (1985). Metacognitive and learning disabilities. In D.L. Forrest-Pressley, G.E. MacKinnon & T.G. Waller (Eds.), *Metacognition, cognition and human performance: Vol. 2. Instructional practices* (pp. 137–180). New York: Academic Press.

Zec, R.F., Parks, R.W., Gambach, J., & Vicari, S. (1992). The executive board system: An innovative approach to cognitive–behavioral rehabilitation in patients with traumatic brain injury. In C.J. Long & L.K. Russ (Eds.), *Handbook of head trauma: Acute care to recovery* (pp. 219–230). New York: Plenum Press.