Journal of the International Neuropsychological Society (2007), **13**, 143–152.

DOI: 10.1017/S1355617707070178

REHAB SERIES

Cognitive rehabilitation in the elderly: Effects on strategic behavior in relation to goal management

BRIAN LEVINE,[1–3] DONALD T. STUSS,[1–4] GORDON WINOCUR,[1,2,5,6] MALCOLM A. BINNS,[1] LOUISE FAHY,[1] MARINA MANDIC,[1] KRISTEN BRIDGES,[1] AND IAN H. ROBERTSON[1,7]

[1]Rotman Research Institute, Baycrest, Toronto, Ontario, Canada
[2]Department of Psychology, University of Toronto, Ontario, Canada
[3]Department of Medicine (Neurology), University of Toronto, Ontario, Canada
[4]Department of Medicine (Rehabilitation Sciences), University of Toronto, Ontario, Canada
[5]Department of Medicine (Psychiatry), University of Toronto, Ontario, Canada
[6]Department of Psychology, Trent University, Peterborough, Canada
[7]Department of Psychology and Institute of Neuroscience, Trinity College, Dublin, Ireland

(RECEIVED November 11, 2005; FINAL REVISION August 4, 2006; ACCEPTED August 8, 2006)

Abstract

Executive functions are highly sensitive to the effects of aging and other conditions affecting frontal lobe function. Yet there are few validated interventions specifically designed to address executive functions, and, to our knowledge, none validated in a healthy aging sample. As part of a large-scale cognitive rehabilitation randomized trial in 49 healthy older adults, a modified Goal Management Training program was included to address the real-life deficits caused by executive dysfunction. This program emphasized periodic suspension of ongoing activity to establish goal hierarchies and monitor behavioral output. Tabletop simulated real-life tasks (SRLTs) were developed to measure the processes targeted by this intervention. Participants were randomized to two groups, one of which received the intervention immediately and the other of which was wait-listed prior to rehabilitation. Results indicated improvements in SRLT performance and self-rated executive deficits coinciding with the training in both groups. These gains were maintained at long-term follow-up. Future research will assess the specificity of these effects in patient groups (*JINS*, 2007, *13*, 143–152.)

Keywords: Neuropsychology, Geriatric assessment, Aging, Frontal lobe, Short-term memory, Intention

INTRODUCTION

The frontal lobes, and connecting systems, are responsible for high-level cognitive operations necessary for the control and direction of lower-level operations, including planning, foresight, multitasking, and self-regulation (Stuss & Levine, 2002; Tranel et al., 1994). Prefrontal damage that causes dysfunction in these higher-level control processes is highly prevalent due to strokes, tumors, dementia, traumatic brain injury, and psychiatric disorders. Because these functions are also affected by damage to pathways, subcortical nuclei, and cortical regions outside the prefrontal cortex, dysfunction in these control processes is among the most prevalent symptoms in patients with brain disease.

While healthy older adults do not have brain damage *per se*, substantial evidence indicates that aging selectively affects the prefrontal cortex (Raz, 2000). Even in the case of general effects of aging on structural and functional brain measures (Greenwood, 2000), processes associated with the prefrontal cortex are still expected to be selectively affected by aging due to prefrontal interconnectivity, as noted above. Thus, aging is known to affect inhibition (Zacks et al., 2000), strategic mnemonic operations (Craik & Grady, 2002; Hashtroudi et al., 1989; Levine et al., 1997, 2002), and concept formation (Kausler, 1991; Levine et al., 1995), among other higher-level functions. This paper is concerned with real-life effects of age-related deficits in higher-level processes, such as forgetting to do things, difficulty solving problems,

Correspondence and reprint requests to: Brian Levine, Ph.D., Rotman Research Institute, Baycrest, 3560 Bathurst St., Toronto, ON M6A 2E1, Canada. E-mail: blevine@rotman-baycrest.on.ca

and disorganization. Rather than the theoretically laden terms of prefrontal or executive function, we will refer to these more generically as *strategic* functions.

In spite of the prevalence of strategic deficits, there are few validated interventions specifically addressing these impairments. In a systematic review of this literature, Turner and Levine (2004; see also Cicerone et al., 2000) identified 40 studies on executive functioning interventions in patients with brain disease, 70% of which were case studies. Only 7% of studies included randomized control groups. Only 48% of studies assessed generalization, with 17% examining real-life outcome. Less than half (48%) included long-term follow-up testing. To our knowledge, there are no standardized, validated protocols for improvement of age-related executive decline.

Although symptom-oriented, pragmatic interventions are often effective, interventions derived from theory may be most likely to produce consistent effects (Green et al., 2004; Robertson & Murre, 1999). Robertson (1996) developed Goal Management Training (GMT), a staged program derived from theories of prefrontal cortex function, most prominently Duncan's (1995) theory of goal neglect.

In GMT, patients are trained to "stop and think" about problems and goals before and during task execution. Although GMT is educational, the emphasis is in interactivity rather than lecturing. Participants complete exercises during the training, perform homework assignments, and use examples from their own lives to illustrate concepts. Training includes identification of absent-minded slips, an appreciation of when these are likely to occur and their consequences. Participants practice checking their "mental blackboards" (i.e., working memory) to ensure that their behavioral output matches their intentions. GMT also includes strategies for dealing with complex, unwieldy tasks (e.g., deciding whether or not to take a new job).

Levine and colleagues (2000) reported an experimental probe of GMT in patients with traumatic brain injury. The patients showed significant improvements on paper and pencil tasks designed to simulate real-life tasks (e.g., proofreading). However, the intervention was very brief (one session), designed as proof-of-principle rather than a *bona fide* rehabilitation. A supplementary case study successfully employed an expanded, eight-session GMT in a postencephalitic patient who sought to improve her cooking ability (Levine et al., 2000). Van Hooren and colleagues (submitted) studied 67 healthy older adults with self-reported strategic deficits and objective evidence of deficits on the Stroop Color-Word Test. Subjects were randomized to either wait-list or GMT groups. Relative to control groups, those receiving GMT reported significantly reduced annoyance as measured by the Cognitive Failures Questionnaire, as well as reduced strategic complaints on an instrument designed for the study. These gains were maintained at 7-week follow-up testing. GMT was not associated with improvements in the Stroop test.

We report the effects of a cognitive rehabilitation program that included a version of GMT (in addition to memory and psychosocial interventions) on lifelike strategic tasks. This version of GMT was modified to fit the constraints of the overall program. The overall goals were to train participants, when confronted with a task, to stop and think about task demands, define the main task, split complex tasks into subtasks (i.e., "Stop-State-Split"), and monitor their performance. Although our focus is on the processes targeted by GMT, the measures reported in this study were administered before and after the entire rehabilitation program, so that the contribution of the memory and psychosocial interventions cannot be ruled out as contributing to the results.

The effects of our cognitive rehabilitation program were assessed with a large battery of tests, including many standard neuropsychological tests and outcome questionnaires (see Craik et al., 2007; Stuss et al., 2007; Winocur et al., 2007a). In the modified GMT, we targeted real-life task performance, which is often disconnected from performance on neuropsychological tests (Alderman et al., 2003; Levine et al., 1998; Shallice & Burgess, 1991). Moreover, our modified GMT focused on task processes rather than performance scores *per se*. Although certain strategic task processes may be inferred from analysis of performance scores (Burgess et al., 1998), it is well known that test scores can be achieved through multiple processes, some advantageous, others not. To distinguish among these processes, we developed novel simulated real-life tasks (SRLTs) from which process ratings were derived through observation and coding of behavior during task performance (Goel et al., 1997). These are paper-and-pencil tasks that resemble complex real-life tasks that are problematic for patients with strategic deficits.

As our intervention trial included multiple assessment probes, we required repeatable tasks, posing a serious problem for assessment of certain strategic processes. Once a strategic task has been administered to a patient, it may no longer be a measure of strategic processes for that patient. To deal with this issue, we created a common underlying structure for our SRLTs and then superimposed different real-life activities over this structure. We report reliability and validity data for one of the SRLTs. We also examined rehabilitation effects using a self-report measure of strategic deficits. Using a crossover design, we assessed a wait-list control group on these same measures, followed by re-assessment after this group received the intervention.

METHODS

Participants and Design

Forty-nine community-dwelling adults 71–87 years of age were recruited from advertisements and word of mouth. For inclusion into the study, participants were required to have subjective complaints of cognitive or memory impairment; they were otherwise in good health. Participants were quasirandomly assigned to an Early Training Group (ETG; $N = 29$) and a Late Training Group (LTG; $N = 20$) with the

https://doi.org/10.1017/S1355617707070178 Published online by Cambridge University Press

constraints that the groups were matched on background variables. Rehabilitation took place immediately after admission into the program for the ETG and 3 months after for the LTG (see Stuss et al., 2007, for more details on the sample). The study was approved by the Baycrest Research Ethics Board and conducted in accordance with the guidelines of the Helsinki Declaration.

This study reports data on 46 participants; data for the other 3 were unavailable due to technical reasons. The ETG had 26 participants (13 men); the LTG had 20 participants (8 men). Missed attendance was negligible. When it occurred, participants were able to make up the session by attending another group or through individual contact with the trainer. The two groups were well matched for mean age (for both groups: 79 years; $SD = 2.4$ and 4.6 for the ETG and LTG, respectively) and mean years education (14.1 and 14.6; $SD =$ 3.2 and 4.7). Neuropsychological test data from a brief set of tests are reported in the Introductory paper to this series (Stuss et al., 2007). The groups were well matched for neuropsychological test results, with one exception being significantly better performance on the WMS-III logical memory subtest in the ETG.

Cognitive Rehabilitation Program

In addition to the modified GMT, the cognitive rehabilitation program included Memory Skills Training and Psychosocial Training modules. Each module lasted 4 weeks. Memory Skills Training (Craik et al., 2007) emphasized the nature of memory loss and the types of aids that could be applied to the process of acquiring, retaining, and recovering information. Psychosocial Training (Winocur et al., 2007a) aimed to enhance psychological well-being and establish the link between overall functional status and cognitive function. The group leader had one-on-one meetings with each participant at the beginning of the trial and on two subsequent occasions. These meetings were intended to set individual goals, answer questions, and address any issues that arose over the course of the trial.

Modified GMT Protocol

The sessions were conducted in an interactive format in which participants generated examples from their own life. Constructs were illustrated with scenarios designed to relate to participants' own life situations. A substantial portion of the sessions consisted of paper and pencil or desktop exercises to promote active application of the constructs, followed by discussion. Sessions 2–4 began with a review of the previous session and discussion of homework.

In the first session, participants were introduced to the construct of absentminded slips, using fictionalized and real-life examples from the trainer and from members of the group. They discussed the emotional and pragmatic consequences of absentminded slips, supplemented by historical examples from industrial accidents. The construct of working memory (the "mental blackboard") was introduced, along with operational definitions of goals and subgoals. Stopping, defined as periodic suspension of ongoing activities to assess goal attainment, was given primary emphasis throughout the session. Participants selected their own catchphrase that they would use as a cue to stop, and this was integrated into previously generated examples in order to illustrate how slips may have been averted by stopping. Automatic pilot errors were defined as absentminded slips due to inappropriate habitual responding. Participants were trained in a simple "present-mindedness" relaxation technique designed to reduce distraction after stopping. The homework assignment for the week following Session 1 consisted of practicing using the catchphrase to stop and think and to log absentminded slips, their antecedents, their consequences, and ideas as to how they might have been prevented.

Session 2 focused on the process of stating or defining the main features of the task at hand. A fictional scenario described an individual who failed to identify a main goal. Techniques for evaluating and prioritizing conflicting goals were described. These emphasized the use of the "Stop!" catchphrase and checking the mental blackboard. A series of examples was presented in which participants identified the conflicting goals and discussed how these might be evaluated (e.g., "You have your fifteen-year high school reunion on the same date as an out of town wedding of a close family friend.") Participants completed and discussed an SRLT involving conflicting goals (this SRLT was not included in the assessment batteries). For homework, participants continued to log absentminded slips. They also completed a log of situations with conflicting goals, including the outcome and evaluative strategies.

In the third session, participants listed complex tasks that make one feel overwhelmed, such as reading furniture assembly instructions. Participants discussed the emotional and practical consequences of these situations. Splitting the task into subtasks was introduced as a management strategy. Abstract problems, involving locating items on a grid were used to illustrate this process, followed by practice with real-life examples, such as preparing a meal. A second SRLT (also separate from those used in the assessment batteries) was completed and discussed. The homework assignments included deconstructing line drawings of variable complexity into a set of verbal instructions and writing instructions to do real-life tasks, such as changing a tire.

In the final session, participants practiced the sequencing and prioritization of subgoals. Again, fictional examples were used to illustrate the need for attention to this matter. Finally, they were introduced to the concept of checking or monitoring, where stopping is used throughout tasks to ensure that outcomes match intentions. The remainder of the session was devoted to discussion and review.

Measures

An extensive battery of outcome measures was administered, including measures of memory and psychosocial out-

https://doi.org/10.1017/S1355617707070178 Published online by Cambridge University Press

come. Data from these measures are reported in the other papers in this volume (Craik et al., 2007; Stuss et al., 2007; Winocur et al., 2007a). In this study, we focus on two measures sensitive to the capacities targeted by GMT: simulated real-life tasks (SRLTs) and the Dysexecutive Questionnaire (DEX; Burgess et al., 1998).

Simulated real life tasks (SRLTs)

The SRLTs were designed to mimic everyday activities that present problems for patients with brain dysfunction due to demands on working memory, attention, and strategic processes. Each SRLT had one main goal that must be achieved by properly arranging several different dimensions or subgoals. Broadly speaking, the SRLTs involved sorting people or objects into groups based on various constraints. Two tasks involved setting up a carpool, one for a school and another for a hospital. The task dimensions included the carpool shift (i.e., morning or afternoon), whether or not the person is a driver or a passenger, available seating, and map location. Participants were provided with a list of 12 people, a map, an answer sheet, and instructions. The list indicated the person's name, whether the person was a driver or passenger, how many passengers the person could take (for drivers), what shift they were assigned to, and their location on the map. The participant was instructed to assign passengers to drivers according to the shift, in the most efficient manner possible considering each passenger's location on the map. A third task, assigning people to swimming lessons, had a parallel structure. In this case, the dimensions were the time of the lesson, the time of swimmers' availability and gender. The tasks contained a "garden path" designed to sidetrack the participant from completing a main goal (e.g., one driver had no passengers to pick up; one swimmer unavailable at the right time). Task instructions also contained irrelevant information, such as general information about the school, requiring participants to extract the relevant information for the task.

Participants were instructed to verbalize their thoughts while performing the SRLT, following the method described by Goel et al. (1997). Before each task, participants solved simple arithmetic problems aloud as a means of practicing this method.

Scoring was done from videotape. Scorers were blind to group membership. For the purposes of scoring, the test session was divided into two phases: (1) reading instructions and preparation and (2) task performance. Participants were rated on four variables: orientation, task strategy, engagement, and checking/error correction. A checklist of six behaviors was developed for each variable within each of the task phases to which it applied (see Table 1). Scorers were blind to group membership. SRLTs were administered in a fixed order. The school carpool task was administered at Assessment A, the hospital carpool task was administered at Assessment B, and the swimming lesson task was administered at Assessment D.

The first draft of the scoring manual consisted of listing and operationalizing elements of GMT that could be observed in participants' SRLT performance. This procedure was then beta-tested by 3 raters using 23 videos randomly selected from the pool, resulting in editing of sections of the manual that proved to be ambiguous. A formal reliability study was then conducted using 10 additional videos and the same 3 raters. The videos used in this study were separate from those used during the development of the manual. Interrater reliability was assessed with the intraclass correlation coefficient (ICC; Shrout & Fleiss, 1979) using a two-way random effects model. The average measure ICCs (appropriate when a subject's protocol is to be rated by multiple raters; Shrout & Fleiss, 1979) were .77, .72, .74, .67, and .84 for orientation, task strategy, engagement, checking, and total SRLT score, respectively. For reference, the cutoff for "excellent agreement beyond chance" is .75 (Fleiss, 1981, p. 218). These data, which reflect high internal consistency reliability as the average ICC measure is very similar to Cronbach's alpha (Shrout & Fleiss, 1979). The single measure ICCs (appropriate when a subject's protocol is to be rated by a single rater; Shrout & Fleiss, 1979) are lower by comparison: .53, .46, .49, .41, and .63 for orientation, task strategy, engagement, checking, and total, respectively. These values are in the range of "fair to good agreement beyond chance" (.40–.75; Fleiss, 1981, p. 218). The SLRT data in the present study were scored by 1 of the 3 trained raters.

Construct validity was assessed by comparing SRLT scores to scores from the Wisconsin Card Sorting Test (WCST; Grant & Berg, 1948), a widely used test of executive functioning. This was done using data from a single SRLT, the school carpool, administered in the initial prerehabilitation assessment (Assessment A, see below) as performance on the remaining SRLTs would be contaminated by the intervention. WCST measures included perseverations to the preceding criterion and loss of set after three or more correct sorts (see Stuss et al., 2000, for definitions and formulas). While there were no significant relationships to WCST perseverative errors, both the checking and total SRLT scores were related to WCST set loss errors, r's ($N = 44$) = .40 and .32, p's $< .01$ and .05, respectively, supporting the validity of the SRLT.

Dysexecutive questionnaire (DEX)

The self-rated form of the DEX (Burgess et al., 1998) was used as a measure of real-life strategic deficits. In contrast to patient populations, healthy adults' self-ratings of real-life reflect a higher symptom endorsement than do others' ratings (Burgess et al., 1998). Because of its relevance to executive function in a social context, the DEX was also included as part of the Psychosocial test battery and results for this test are also reported in the paper by Winocur et al. (2007a).

Procedure

The ETG was assessed at pretraining, posttraining, and long-term follow-up. The LTG was assessed at pretraining and

https://doi.org/10.1017/S1355617707070178 Published online by Cambridge University Press

Table 1. Simulated real-life task scoring guide

Scoring category	Task phase	
	Instructions	Performance
Orientation	1. Looking at all relevant material 2. Pacing self with hands or pencil while reading through instructions 3. Underlining, taking notes or otherwise marking important points in instructions 4. Cross-referencing all supplementary materials 5. Looking at answer sheet in preliminary attempt to organize strategy 6. Reorganizing material for own ease of use	
Task strategy		1. Systematic approach to task 2. Prioritizing relevant dimensions for easiest task completion 3. Recognizing all relevant dimensions 4. Recognizing garden path 5. Appropriate amount of time spent on relevant dimensions 6. Persistence in completing all dimensions
Engagement		1. Leaning in 2. Pointing at important points, highlighting materials 3. Positive facial expression 4. Positive tone of voice 5. Positive verbal expressions 6. Gaze focused on task
Monitoring, Error correction	1. Translation of main idea into own idiom, not reciting material from instructions 2. Recognizing 3/4 relevant dimensions, not focusing on irrelevant details as though necessary for task completion 3. Evidence of plan to deal with task dimensions making deductions about task based on dimensions 4. Evidence of plan to deal with garden paths 5. Looking for own solutions, not asking questions of administrator about task-related material 6. Counting elements in dimensions	1. Translation of dimensions into own idiom, not reciting from instructions 2. Taking control over all supplementary materials by writing own codes/symbols on them 3. Progressing through at least 50% of dimensions 4. Looking for own solutions when faced with incongruity in materials, not asking questions of administrator about task-related material 5. Restatement at key points of current position in goal hierarchy 6. Checking correctness of responses either verbally or non-verbally

long-term follow-up. For both groups, long-term follow-up occurred 6 months following the last session. To assess the effects of time and test interval, LTG participants had a supplementary pretraining assessment coinciding with the ETG's pretraining assessment. The first assessment for both groups (pretraining for ETG and supplementary pretraining for LTG) will be referred to as Assessment A. The second assessment (posttraining for the ETG, pretraining for the LTG) will be referred to as Assessment B. Assessment C took place immediately posttraining for the LTG. For technical reasons, there were no SRLTs administered at Assessment C. The long-term follow-up assessment will be referred to as Assessment D. As indicated in the Introductory paper (Stuss et al., 2007), the sample size in the LTG was diminished for Assessment D (and to some degree C) compared to the other assessments due to the rapid turn-over of personnel toward the end of the trial.

Statistical Analysis

The statistical analysis strategy followed the general framework as described by Stuss et al. (2007), with a few exceptions. The main comparison of interest was at Assessment B, where the two groups were matched for time enrolled in the study and exposure to the outcome measures (i.e., both groups had received the school carpool task and the DEX 14 weeks earlier). For this analysis, data were available for 22 ETG participants and 16 LTG participants. As indicated in the Introductory paper, 5 participants in the LTG did not complete Assessments C and D due to an outbreak of severe acute respiratory syndrome (SARS), which effectively prohibited research participants from entering the hospital. Data were analyzed by means of analysis of covariance, directly comparing performance on the hospital carpool task across groups, statistically controlling for performance on the school

https://doi.org/10.1017/S1355617707070178 Published online by Cambridge University Press

carpool task. Supplementary analyses of variance (ANOVAs) were conducted to examine between- and within-group effects across Assessments A and B. The η^2 (eta-squared) statistic is used to report effect size. According to Cohen (1988), thresholds for interpreting η^2 are less than .06 (small), .06 to .14 (medium), and greater than .14 (large).

Maintenance of effects over time was assessed by examining SRLT performance at Assessment D (i.e., on the swimming lesson task at long-term follow-up) as compared with Assessment A. Data from 16 ETG and 12 LTG participants were available for this analysis. Data were analyzed with repeated-measures ANOVAs separately for each group, whereby performance at long-term follow-up was directly compared with the initial school carpool task. Assessment D was conducted 6 months posttraining for both groups. As another perspective on the long-term follow-up data, the ETG and LTG were compared directly using ANOVA at Assessment D, with the expectation that the effects of the intervention would reduce between-group differences apparent at Assessment B. Analysis of the DEX data paralleled that of the SRLT data, except that data were available for more subjects: 47 at Assessments A and B and 44 at Assessment D.

RESULTS

Simulated Real-Life Tasks (SRLTs)

Rehabilitation effects on SRLT performance are depicted in Table 2 and Figure 1. With a few exceptions, the pattern across measures is that ETG and LTG participants performed similarly at Assessment A, diverged at Assessment B with the ETG improving their scores following the intervention, then converged at Assessment D where both scores stabilized at the higher level.

Table 2. SRLT performance

		Group			
		ETG		LTG	
Task	Assessment	*M* (*SD*)	Range	*M* (*SD*)	Range
Checking	A	5.96 (2.52)	0–9	5.95 (2.28)	2–10
	B	7.88 (1.94)	3–11	5.81 (2.54)	2–11
	D	8.78 (1.66)	5–11	7.50 (1.93)	4–10
Engagement	A	8.23 (2.66)	4–12	7.25 (2.59)	3–12
	B	9.08 (2.12)	5–12	7.13 (2.66)	2–11
	D	9.94 (2.13)	5–12	8.75 (2.30)	5–12
Orientation	A	3.35 (1.55)	0–5	3.05 (1.43)	1–6
	B	3.08 (1.68)	0–6	2.44 (1.41)	1–5
	D	3.28 (1.67)	0–6	2.69 (1.55)	0–4
Task Strategy	A	2.58 (1.70)	1–6	2.55 (1.19)	0–5
	B	4.36 (1.38)	2–6	2.50 (1.55)	0–6
	D	4.44 (1.50)	2–6	3.33 (1.92)	0–6
Total	A	16.12 (6.05)	4–25	15.45 (5.02)	6–24
	B	20.20 (4.86)	12–28	14.44 (5.05)	4–24
	D	21.28 (4.17)	13–28	18.00 (4.88)	8–25

Note. ETG = Early Training Group; LTG = Late Training Group.

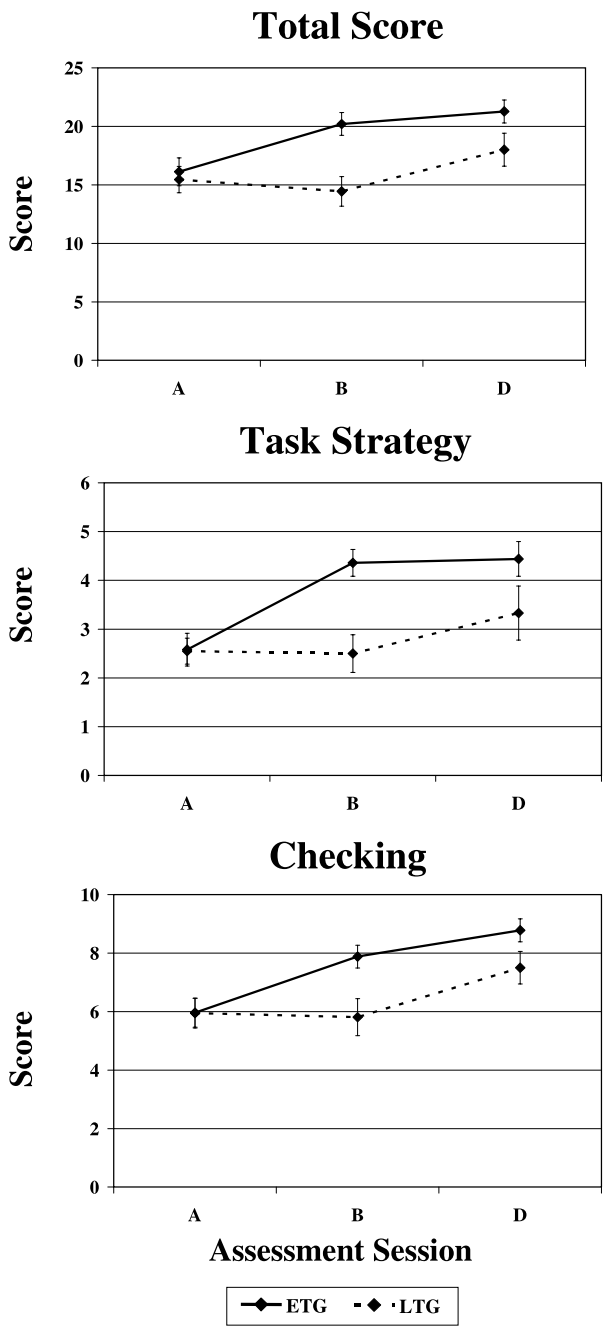

Fig. 1. Effects of rehabilitation on simulated real-life task (SRLT) performance. Each panel represents a different SRLT score (see Table 1 for explanation). For all three scores, the Early Training Group (ETG) improved immediately following rehabilitation (Assessment B), with improvements maintained at follow-up (Assessment D). The Late Training Group (LTG) remained stable over the first two prerehabilitation sessions (Assessments A and B), and showed gains at follow-up (Assessment D). Error bars represent the standard error of the mean.

At Assessment B (statistically controlling for performance at Assessment A), there was a significant effect of group on the overall GMT score [$F(1,35) = 10.95$, $p = .002$, $\eta^2 = .24$], with the ETG scoring higher than the LTG. Similar effects were noted for task strategy [$F(1,35) = 15.14$,

https://doi.org/10.1017/S1355617707070178 Published online by Cambridge University Press

$p < .001$, $\eta^2 = .30$], checking [$F(1,35) = 8.21$, $p = .007$, $\eta^2 = .19$]; and engagement [$F(1,35) = 4.77$, $p = .04$, $\eta^2 = .12$]. These results could not be accounted for by statistical artifact stemming from the use of the covariate; ETG, but not LTG, participants showed significant improvement from Assessment A to Assessment B for total score, task strategy, and checking [F's(1,21) = 6.56, 18.05 and 10.41; p's = .02, .001, and .004; and $\eta^2 = .24$, .46, and .33, respectively, for ETG]. The ETG did not significantly improve on the engagement variable from Assessment A to B. There were no significant effects involving the orientation score.

At Assessment D, ETG participants maintained increased scores relative to Assessment A for total score [$F(1,15) = 14.49$, $p = .002$, $\eta^2 = .49$], task strategy [$F(1,15) = 15.63$, $p = .001$, $\eta^2 = .51$], and checking [$F(1,15) = 31.12$, $p < .001$, $\eta^2 = .68$]. The effect for engagement fell outside the range of significance [$F(1,15) = 3.38$, $p = .09$, $\eta^2 = .18$]. The LTG, who did not improve at Assessment B, now showed gains at Assessment D for total score [$F(1,11) = 5.32$, $p = .04$, $\eta^2 = .33$] and checking [$F(1,11) = 9.52$, $p = .01$, $\eta^2 = .46$], but task strategy, engagement, or orientation were not significant. There were no significant between-group differences at Assessment D. As seen in Figure 1, however, the ETG scored nonsignificantly higher than the LTG for total score, task strategy, and checking at Assessment D [$F(1,28) = 3.89$, $p = .06$, $\eta^2 = .12$; $F(1,28) = 3.15$, $p = .09$, $\eta^2 = .10$; and $F(1,28) = 3.74$, $p = .06$, $\eta^2 = .12$, respectively].

Dysexecutive Questionnaire (DEX)

Data for the DEX are presented in Table 3. At Assessment B, the group effect for the DEX, statistically controlling for Assessment A scores, fell outside the range of significance [$F(1,45) = 2.06$, $p = .16$, $\eta^2 = .04$], although endorsement of executive deficits for the ETG was nonsignificantly lower in Assessment B as compared to Assessment A [$F(1,27) = 3.03$, $p = .09$, $\eta^2 = .10$]. There was a significant effect at Assessment C controlling for A [$F(1,26) = 12.70$, $p = .001$, $\eta^2 = .33$]. The LTG showed no significant change in DEX scores between Assessments B and C, although there was a significant effect at Assessment C controlling for A [$F(1,14) = 15.61$, $p = .001$, $\eta^2 = .53$]. Both groups endorsed fewer items on the DEX at Assessment D as compared to Assessment A [for ETG, $F(1,27) = 10.16$, p = .004, $\eta^2 = .27$; for LTG, $F(1,14) = 9.15$, $p = .01$, $\eta^2 = .40$]. There were no significant between group differences for DEX scores at Assessment D.

Table 3. DEX performance

	Group			
	ETG		LTG	
Assessment	*M* (*SD*)	Range	*M* (*SD*)	Range
A	15.45 (8.97)	1–30	17.80 (8.21)	7–35
B	13.21 (7.17)	2–31	16.90 (7.40)	4–30
C	11.59 (7.16)	0–24	14.33 (7.59)	3–29
D	12.79 (8.45)	0–29	13.73 (7.54)	3–32

Note. ETG = Early Training Group; LTG = Late Training Group.

DISCUSSION

Strategic processes associated with prefrontal function are sensitive to a wide variety of brain changes, including those accompanying normal aging, with significant implications for quality of life. Goal Management Training (GMT; Robertson, 1996) is a novel, standardized intervention that attempts to improve goal attainment through training simple attentional and organizational skills. A modified version of GMT was incorporated into a comprehensive cognitive rehabilitation program for older adults (Stuss et al., 2007). We found that older adults significantly improved their performance on SRLTs and reduced self-reported executive failures following training. The training effects appeared to be specific to the rehabilitation intervention, inasmuch as they were not observed in the wait-list control (LTG) group until training was applied. Furthermore, gains were maintained at long-term follow-up assessment.

Assessing Rehabilitation Effects With SRLTs

The SRLTs were specifically designed for the present study, as there are no widely used, validated measures of process-based assessment of the constructs targeted by GMT. Because of the novelty of these instruments, we conducted a separate study to examine their reliability and validity. Overall, our data support the reliability of the SRLTs and their validity as strategic measures. The reliability study indicated intra-class correlation coefficients (ICCs) at or near the "excellent" level for the school carpool SRLT when scores are averaged across multiple raters. This measure also reflects internal consistency reliability (i.e., coefficient alpha; Shrout & Fleiss, 1979). In most applications (including the present rehabilitation study), protocols are scored by only 1 rater, in which case the lower reliabilities for the single measure ICC should apply. These reliability coefficients, although lower, are still within an acceptable range. To the extent that the reliability of the SRLT is mildly compromised, this would work *against* our findings by increasing the noise in the data. That is, the true effect sizes of this intervention on the underlying construct assessed by the SRLT are likely larger than we were able to detect.

There is no gold standard against which to assess the validity of the SRLTs. The WCST, however, is a widely used measure of executive functioning that has some overlap with the constructs assessed by the SRLT. SRLT checking and, to a lesser extent, the SRLT total score were negatively correlated with WCST set loss errors, which can be interpreted as a measure of checking or monitoring ongoing responses (Stuss et al., 2000). WCST perseverative errors, on the other hand, reflect a different construct (incorpora-

https://doi.org/10.1017/S1355617707070178 Published online by Cambridge University Press

tion of examiner feedback) and were not significantly related to SRLT performance. These data suggest that the SRLT may be well suited as an outcome measure for interventions that target "stop and think" behaviors, like GMT. Nonetheless, given the absence of widely accepted, standardized measures of real-life functioning, the generalizability of our findings to behavior outside the laboratory could not be assessed.

The three SRLTs in our battery share the same underlying structure, but with a different 'interface' designed to reduce practice effects that can affect interpretation of executive functioning tasks. Our reliability and validity study was restricted to a single SRLT (from Assessment A), as this was the only one administered prior to rehabilitation. While nonequivalence of tasks cannot be ruled out, it is suggested that it is the underlying structure, rather than the superficial user interface, that is critical in the SRLTs. In any case, the potential confound of nonequivalence does not affect interpretation of dissociations across groups on the same tasks, as was the case of Assessment B, where the ETG reliably improved, but the LTG did not. In the instance where *both* groups improved (as was the case for Assessment D), however, it is impossible to separate task effects from group effects. In other words, differences in task difficulty cannot be ruled out as contributing to improved performance at Assessment D.

Patterns of Change in Test Scores

Improvement in SRLT scores was associated with the intervention; their elevation remained stable at two posttraining assessments in the ETG, while the baseline level was stable at two pretraining assessments in the LTG. The total SRLT score, a composite of the other scores, attained the highest interrater reliability, at the upper end of the "fair to good range" (Shrout & Fleiss, 1979). This score was sensitive to intervention effects at every test interval for both groups. Of the SRLT process scores, checking was the most closely related to the content of the modified GMT and was additionally the most sensitive to intervention effects. Task strategy also showed consistent effects. Orientation and engagement were the least sensitive, possibly due to the more subjective nature of scoring criteria for these process scores.

Although the LTG showed treatment gains following rehabilitation, their long-term follow-up scores (Assessment D) did not reach the levels of the ETG, in spite of the fact that the LTG long-term follow-up assessment occurred more proximal to the intervention. These group differences fell short of statistical significance, but power was also reduced at this assessment interval. As noted elsewhere (Winocur et al., 2007a,b), the LTG was at a consistent disadvantage relative to the ETG, possibly due to the effects of delayed intervention.

The DEX is a widely used self- and significant-other report of real-life executive deficits. In patients with brain disease, other-rated DEX scores are more sensitive than are self-rated scores, presumably due to insight deficits in patients (Burgess et al., 1998). This is not the case, however, for healthy adults, whose self-rated scores are higher than their other-rated scores (Burgess et al., 1998). We, therefore, adopted DEX self-ratings for this study. The DEX results followed a somewhat different pattern. Neither group showed significant improvement on the DEX immediately posttraining, although the ETG improved marginally. Both groups, however, showed significantly lower DEX scores at long-term follow-up. One possibility is that the real-life effects assessed by the DEX required time to consolidate. The postrehabilitation assessment may have been too early to detect effects, but by the time of long-term follow-up, participants had integrated training effects into their lives. However, because both groups improved, the effects of multiple DEX administrations cannot be ruled out as contributing to these effects. Furthermore, because the DEX is a self-report instrument, demand characteristics may have influenced the results.

Limitations and Caveats

We included two SRLT tasks as part of the intervention. Although these tasks were not used in the assessments, it is possible that practice on these tasks may have contributed to SRLT intervention effects. It is unlikely, however, that the positive outcomes can be fully accounted for by the inclusion of SRLTs in the training. The SRLTs comprised a relatively small proportion of the training. Furthermore, SRLT effects were observed for measures specific to processes targeted by the training (monitoring, task strategy) that would be unlikely to obtain from mere practice on the SRLTs without instruction.

The LTG did not receive a posttraining assessment with an SRLT immediately following the intervention, as did the ETG. We were, therefore, unable to properly assess maintenance of gains for this group. Nonetheless, given that the ETG maintained their gains and that the LTG remained stable over the first two assessments, it appears likely that LTG's improvement at long-term was due to the intervention.

This group of healthy elderly subjects did not have frank neuropsychological impairments, although subjective complaints of cognitive or memory changes were required for inclusion in this study. The overall intact cognitive status of this sample set a limit on the possible rehabilitation gains, which were modest, even though effect sizes ranged from moderate to large. Further research in more impaired populations is needed to assess the practical significance of these rehabilitation effects.

The outcome measures reported in this paper were designed to be sensitive to the effects of modified GMT. However, it is not possible to draw any conclusions about the specificity of this effect to the modified GMT portion of our rehabilitation intervention program as assessments were taken before and after the entire program, which also included psychosocial and memory training. Additionally,

although this trial was randomized, nonspecific effects of the intervention (e.g., contact with the group and trainer) cannot be ruled out as explaining the results. We are currently addressing these issues with an expanded, updated GMT intervention in patients with brain injury using a randomized controlled trial design in which both groups receive a viable intervention rather than comparisons being made to a placebo condition.

CONCLUSIONS

Healthy older adults significantly improved on both examiner-rated performance laboratory analogues of real-life tasks and on self-rated executive functioning following administration of a cognitive rehabilitation intervention that included a modified Goal Management Training protocol designed to increase real-life goal attainment through interactive, task-based training in attentional control and self-organization. Executive deficits pose significant functional disability in both aging and brain damaged populations. Future research will assess the specificity of these findings relative to other interventions and to patients with brain disease.

ACKNOWLEDGMENTS

The authors acknowledge the valuable contributions of the other co-investigators on this project: Drs. M. Alexander, S. Black, F.I.M. Craik, and D. Dawson. Vinod Goel is thanked for advice and materials concerning the verbal monitoring methodology. As well, the outstanding support provided by Rayonne Chavannes, Maureen Downey-Lamb, Marie-Ève Couture, Melissa Edwards, Peter Glazer, Tara McHugh, and Heather Palmer is gratefully acknowledged. This study, and the experimental trial of which it is a part, were supported by the JSF McDonnell Foundation. D. Stuss was supported by the Reva James Leeds Chair in Neuroscience and Research Leadership. The information reported in this manuscript and the manuscript itself are new and original. The manuscript is not under review by any other journal and has never been published either electronically or in print. There are no financial or other relationships that could be interpreted as a conflict of interest affecting this manuscript.

REFERENCES

Alderman, N., Burgess, P.W., Knight, C., & Henman, C. (2003). Ecological validity of a simplified version of the multiple errands shopping test. *Journal of the International Neuropsychological Society*, *9*, 31–44.

Burgess, P.W., Alderman, N., Evans, J., Emslie, H., & Wilson, B.A. (1998). The ecological validity of tests of executive function. *Journal of the International Neuropsychological Society*, *4*, 547–558.

Cicerone, K.D., Dahlberg, C., Kalmar, L., Langenbahn, D., Malec, J.F., Bergquist, T.F., Felicetti, T., Giacino, J.T., Harley, J.P., Harrington, D.E., Herzog, J., Kneipp, S., Laatsch, L., & Morse, P. (2000). Evidence-based cognitive rehabilitation: Recommendations for clinical practice. *Archives of Physical Medicine and Rehabilitation*, *81*, 1596–1615.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Craik, F.I.M. & Grady, C.L. (2002). Aging, memory, and frontal lobe functioning. In D.T. Stuss & R. Knight (Eds.), *Principles of frontal lobe function* (pp. 528–540). New York: Oxford University Press.

Craik, F.I.M., Winocur, G., Palmer, H., Binns, M.A., Edwards, M., Bridges, K., Glazer, P., Chavannes, R., & Stuss, D.T. (2007, this issue). Cognitive rehabilitation in the elderly: Effects on memory. *Journal of the International Neuropsychological Society*, *13*, 132–142.

Duncan, J. (1995). Attention, intelligence, and the frontal lobes. In M.S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 721–733). Cambridge, MA: MIT press.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*. Washington, DC: John Wiley & Sons, Inc.

Goel, V., Grafman, J., Tajik, J., Gana, S., & Danto, D. (1997). A study of the performance of patients with frontal lobe lesions in a financial planning task. *Brain*, *120*, 1805–1822.

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

Green, R.E., Turner, G.R., & Thompson, W.F. (2004). Deficits in facial emotion perception in adults with recent traumatic brain injury. *Neuropsychologia*, *42*, 133–141.

Greenwood, P.M. (2000). The frontal aging hypothesis evaluated. *Journal of the International Neuropsychological Society*, *6*, 705–726.

Hashtroudi, S., Johnson, M.K., & Chrosniak, L.D. (1989). Aging and source monitoring. *Psychology and Aging*, *4*, 106–112.

Kausler, D.H. (1991). Thinking: Concept formation and identification. In *Experimental psychology, cognition, and human aging* (2nd ed.) (pp. 552–595). New York: Springer-Verlag.

Levine, B., Robertson, I.H., Clare, L., Carter, G., Hong, J., Wilson, B.A., Duncan, J., & Stuss, D.T. (2000). Rehabilitation of executive functioning: An experimental-clinical validation of goal management training. *Journal of the International Neuropsychological Society*, *6*, 299–312.

Levine, B., Stuss, D.T., & Milberg, W.P. (1995). Concept generation: Validation of a test of executive functioning in a normal aging population. *Journal of Clinical and Experimental Neuropsychology*, *17*, 740–758.

Levine, B., Stuss, D.T., & Milberg, W.P. (1997). Effects of aging on conditional associative learning: Process analyses and comparison with focal frontal lesions. *Neuropsychology*, *11*, 367–381.

Levine, B., Stuss, D.T., Milberg, W.P., Alexander, M.P., Schwartz, M., & Macdonald, R. (1998). The effects of focal and diffuse brain damage on strategy application: Evidence from focal lesions, traumatic brain injury, and normal aging. *Journal of the International Neuropsychological Society*, *4*, 247–264.

Levine, B., Svoboda, E., Hay, J., Winocur, G., & Moscovitch, M. (2002). Aging and autobiographical memory: Dissociating episodic from semantic retrieval. *Psychology and Aging*, *17*, 677–689.

Raz, N. (2000). Aging of the brain and its impact on cognitive performance: Integration of structural and functional findings. In F.I.M. Craik & T.A. Salthouse (Eds.), *The handbook of aging and cognition* (2nd ed.) (pp. 1–90). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Robertson, I.H. (1996). *Goal management training: A clinical manual*. Cambridge, UK: PsyConsult.

Robertson, I.H. & Murre, J.M. (1999). Rehabilitation of brain damage: Brain plasticity and principles of guided recovery. *Psychological Bulletin*, *125*, 544–575.

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

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Stuss, D.T. & Levine, B. (2002). Adult clinical neuropsychology: Lessons from studies of the frontal lobes. *Annual Review of Psychology*, *53*, 401–433.

Stuss, D.T., Levine, B., Alexander, M.P., Hong, J., Palumbo, C., Hamer, L., Murphy, K.J., & Izukawa, D. (2000). Wisconsin Card Sorting Test performance in patients with focal frontal and posterior brain damage: Effects of lesion location and test structure on separable cognitive processes. *Neuropsychologia*, *38*, 388–402.

Stuss, D.T., Robertson, I.H., Craik, F.I.M., Levine, B., Alexander, M.P., Black, S., Dawson, D., Binns, M.A., Palmer, H., Downey-Lamb, M., & Winocur G. (2007, this issue). Cognitive rehabilitation in the elderly: A randomized trial to evaluate a new protocol. *Journal of the International Neuropsychological Society*, *13*, 120–131.

Tranel, D., Anderson, S.W., & Benton, A. (1994). Development of the concept of "executive function" and its relationship to the frontal lobes. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. 9) (pp. 125–148). Amsterdam: Elsevier.

Turner, G.R. & Levine, B. (2004). Disorders of executive function and self-awareness. In J. Ponsford (Ed.), *Rehabilitation of neurobehavioral disorders* (pp. 224–268). New York: Guilford Publications.

van Hooren, S.A.H., Valentijn, A.M., Bosma, H., Ponds, R.W.H.M., Boxtel, M.P.J.v., Levine, B., Robertson, I.H., & Jolles, J. (in press). Effect of a structured course involving Goal management training in older adults: A randomized controlled trial. *Patient Education and Counseling*.

Winocur, G., Craik, F.I.M., Levine, B., Robertson, I.H., Binns, M.A., Alexander, M.P., Black, S., Dawson, D., Palmer, H., Downey-Lamb, M., & Stuss, D.T. (2007b, this issue). Cognitive rehabilitation in the elderly: Overview and future directions. *Journal of the International Neuropsychological Society*, *13*, 166–171.

Winocur, G., Palmer, H., Dawson, D., Binns, M.A., Bridges, K., & Stuss, D.T. (2007a, this issue). Cognitive rehabilitation in the elderly: An evaluation of psychosocial factors. *Journal of the International Neuropsychological Society*, *13*, 153–165.

Zacks, R.T., Hasher, L., & Li, D.Z.H. (2000). Human memory. In F.I.M. Craik & T.A. Salthouse (Eds.), *The handbook of aging and cognition* (pp. 293–357). Mahwah, NJ: Lawrence Erlbaum Associates.