Computer- and Suggestion-based Cognitive Rehabilitation following Acquired Brain Injury.

Lindeløv, Jonas Kristoffer

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

COMPUTER- AND SUGGESTION-BASED COGNITIVE REHABILITATION FOLLOWING ACQUIRED BRAIN INJURY

BY
JONAS KRISTOFFER LINDELØV

DISSERTATION SUBMITTED 2015

AALBORG UNIVERSITY
DENMARK

Computer- and Suggestion-based Cognitive Rehabilitation following Acquired Brain Injury

BY
JONAS KRISTOFFER LINDELØV

AALBORG UNIVERSITY
DENMARK

DISSERTATION SUBMITTED 2015

Thesis submitted: April, 2015

PhD supervisor: Professor Morten Overgaard
Aalborg University

PhD committee: Professor Jesper Mogensen
University of Copenhagen

Professor Zoltan Dienes
University of Sussex

Professor Mårten Riesling
Karolinska Institutet

PhD Series: Faculty of Humanities, Aalborg University

ISSN (online): 2246-123X
ISBN (online): 978-87-7112-290-9

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2015

Summary

This thesis is motivated by an unfortunate state of affairs: impairment of working memory and attention is a frequent sequelae of acquired brain injury experienced by millions of people all over the world. However, there are currently no cost-effective treatments for such impairments. The present work is an empirical exploration of two candidate treatments to ameliorate this problem.

Computer-based cognitive rehabilitation is one such candidate treatment although results are mixed in healthy- and patient populations. Interestingly, research on these two populations have proceeded in parallel with little cross-talk. We recruited 39 healthy subjects and 39 inpatients in the post-acute phase. They were stratified to an N-back condition or a closely matched Visual Search condition. Training did not cause transfer to untrained tasks in neither condition for neither subject group. However, healthy subjects improved 2.5 to 5.5 standardized mean differences (SMD) more than patients on the training tasks, indicating an impaired ability to learn task-specific strategies in the patient population.

To further explore the effectiveness of computer-based cognitive training, I conducted a review and meta-analysis of the literature. Thirty studies were included in the review and 26 in the meta-analysis, representing a total of 1086 patients across 42 treatment-control contrasts. There was no transfer effect (SMD=0.05) when controlling for task-specific learning (coded as training-test similarity) and inappropriate control conditions. From this analysis and the N-back experiment, it is concluded that computer-based training in general is unsuitable as a means to rehabilitate high-level cognitive functions in patients with acquired brain injury.

In a third study, we used hypnotic suggestions to target impaired working memory capacity. 68 brain injured patients in the chronic phase were included. There was a unique effect of these targeted hypnotic suggestions as compared to a closely matched hypnotic control condition (SMD=0.66) and very large effects compared to a passive control group (SMD=1.15). This effect was sustained after a 7 week break. The hypnotic control condition improved when crossed over to the active condition, resulting in both groups having normalized performance on working memory outcome measures while a passive control group remained at status quo throughout the experiment.

In light of these results and the broader literature on the topic, I suggest that we should be mindful of our intuitions about how a given intervention affects the mind. I present three metaphors which allow for a quick heuristic reasoning about the expected outcomes of

different families of interventions. In the typical case, high-level cognition is not improved by exercising it bottom-up as a muscle, as is done in computer-based training. Neither is it improved by nourishing it like a plant, as seems to be the strategy of physical exercise and pharmaceuticals. I suggest that the metaphor of the mind as a computer which can be programmed using symbolic language may carry useful intuitions to predict the outcome of the interventions considered in this thesis.

Dansk resume

Denne afhandling er motiveret af et uheldigt sammenfald: nedsat arbejdshukommelses- og opmærksomheds-funktion er en af de hyppigst forekommende følgevirkninger blandt de millioner af mennesker, som har en erhvervet hjerneskade på verdensplan. Alligevel findes der ikke effektive behandlinger af disse følgevirkninger. Denne afhandling er en empirisk undersøgelse af to interventioner, som er kandidater til at afhjæpe problemet.

Computer-baseret kognitiv rehabilitering har hidtil vist blandede resultater i den almene befolkning og med patienter. Forskningen i disse to populationer har dog kørt parallelt uden sammenligninger eller gensid inspiration. Som et første skridt til at bygge bro mellem disse to felter, udførte vi et forsøg hvor vi rekrutterede 39 raske forsøgspersoner og 39 patienter i den post-akutte fase. Deltagerne blev stratificeret til enten at træne N-back opgaven eller en tæt matchet opgave med visuel skanning. Ingen af opgaverne medførte transfer til ikke-trænede tests for nogle af grupperne. Dog forbedrede de raske deltagere sig langt mere på selve træningsopgaven end patienterne, hvilket indikerer en nedsat evne til at tilegne sig opgave-specifikke strategier efter erhvervet hjerneskade.

For yderligere at undersøge effektiviteten af computer-baseret træning på høj-niveau kognition, udførte jeg et review og en meta-analyse. I alt blev 30 artikler inkluderet i reviewet, hvoraf 26 indgår i meta-analysen. Datasættet dækker i alt 1086 patienter fordelt på 42 intervention-kontrol kontraster. Der var ingen transfer effekt (SMD=0.05) når der blev kontrolleret for opgave-specifik læring (lighed mellem træningsopgave og test) og utilstrækkeligt matchede kontrolgrupper. Tilsammen indikerer N-back forsøget og meta-analysen at computer-træning ikke er egnet til rehabilitering af høj-niveau kognition i patienter med erhvervet hjerneskade.

I et tredje forsøg anvendte vi hypnotiske suggestioner om at forbedre arbejdshukommelses-kapaciteten. 68 personer med kroniske følger af erhvervet hjerneskade blev inkluderet. Der var en mellemstor unik effekt af de målrettede suggestioner, sammenlignet med en tæt matchet kontrolgruppe (SMD=0.66) og en meget stor effekt sammenlignet med en passiv kontrolgruppe (SMD=1.15). Denne effekt var opretholdt efter 7 ugers pause. Kontrolgruppen forbedrede sig da de herefter blev krydset over til de målrettede

suggestioner, med det resultat, at begge grupper opnåede en normaliseret ydeevnet på arbejdshukommelsestests. En passiv kontrolgruppe forblev på status quo igennem hele eksperimentet.

På baggrund af disse resultater og den bredere literatur på området, opfordrer jeg til refleksion over vores implicitte intuitioner om hvordan en given intervention kan påvirke psyken. Som afrunding præsenterer jeg tre metaforer, der heuristisk kan give en intuition om de forventede udfald af forskellige typer af interventioner. Høj-niveau kognition forbedres sjældent ved at en stimulus-dreven træning af hjernen som var den en muskel, sådan som det for eksempel gøres med computer-baseret træning. At nære hjernen med psykofarmaka eller motion, som var det en plante, medfører heller ikke forbedringer. Jeg foreslår, at et metafor om psyken som en computer der kan programmeres med et symbolsk sprog, kan indfange nyttige intuitioner, som kan forudsige de eksperimentelle resultater fra denne afhandling.

Table of contents

This thesis has four parts. The background chapter could be considered to be an "experimental toolbox". The chapter on computer-based cognitive rehabilitation contains an empirical paper (henceforth referred to as the "N-back paper") and a review paper. The chapter on suggestion-based rehabilitation contains an empirical paper (henceforth referred to as the "hypnosis paper"). The main discussion takes place in chapter two and three but it is followed up by a short overall discussion in chapter four.

Definitions

High-level cognition: cognitive functions which take a wide variety of stimuli as input across multiple modalities and transform them to behavior in a flexible way. Examples include attention, working memory, and executive functions.

Low-level cognition: cognitive functions which take a narrow set of stimuli as input, perhaps only in one modality and transform them to behavior in a fairly rule-based way. Examples include pitch discrimination, coordination of hand-movements, and naming familiar objects.

Probabilistic inference: assigning probabilities to possible causes (e.g. parameters or parameter values) of an observed output.

Transfer effect: a change in performance on non-trained behaviors.

Training effect: a change in performance on trained behaviors.

Nomenclature

APT: Attention Process Training
BF: Bayes Factor.
HNT: Hours Needed to Treat. HNT = NNT x hours
LRT: Likelihood Ratio Test
MBSR: Mindfulness Based Stress Reduction
NHST: Null Hypothesis Significance Testing
NNT: Number Needed to Treat. In this thesis calculated from SMD.
OSPAN: Operation Span - a complex span task.
RAPM: Raven's Advanced Progressive Matrices.
SMD: Standardized Mean Difference. SMD = $(\mu_{post} - \mu_{pre}) / \square_{baseline}$ where μ is the population mean and $\square$ is the population (i.e. unbiased) standard deviation.
SMD_c: the controlled Standardized Mean Difference. $SMD_c = SMD_{treat} - SMD_{control}$.
TBI: Traumatic brain injury, including concussion.

List of figures

List of tables

Paper publication status

As of submitting this thesis, the N-back paper is in review in Neuropsychological Rehabilitation for a special issue on “brain training”.

The review paper is currently in review in Clinical Psychology Review.

The hypnosis paper will be submitted to Plos Biology shortly.

All data and analysis scripts will be made publicly available at https://osf.io/zxgjn when each paper is accepted for publication. Further materials can be obtained by contacting me on jonas@cnru.dk

1 BACKGROUND

1.1 Brain injury and neuropsychological rehabilitation

1.1.1 Epidemiology

The number of people for whom the topic of this thesis is relevant is staggering. In the U.S. alone, an estimated 3.3 to 5.3 million people suffered from long-term disabilities following traumatic brain injury (TBI) according to a 2005-estimate (Ma, Chan, & Carruthers, 2014). In 2010 there was an estimated 23 million stroke survivors under 75 years of age globally (Feigin et al., 2014). This is an increase in 10 million compared to 1990, likely caused by the improvement in life-saving procedures worldwide. A consequence of more people surviving injuries to the central nervous system is that more people live with disabilities (American Heart Association, 2014). Therefore, chronic sequelae of brain injury constitutes an increasing societal cost of healthcare, assistance with daily living, subsidiaries, and lost productivity. Such costs are estimated to be in the order of an annual 48-76 billion for TBI and 34-65 billion for stroke in the US alone[1]. This makes stroke and TBI the second and third largest health care cost in the US next to pain (Ma et al., 2014). As the population life span increases, the incidence rate of stroke due to cardiovascular degeneration and TBI due to falls are expected to increase even more (Feigin, Lawes, Bennett, & Anderson, 2003).

1.1.2 Sequelae

46% of ischemic stroke survivors suffer from long-term cognitive deficits (Kelly-Hayes et al., 2003). Most reviews qualitatively point to memory and attention as the most frequent impairments without giving specific numbers (Cumming, Marshall, & Lazar 2013; Carroll et al., 2003). Consistent with this, TBI patients primarily complain about forgetfulness and poor concentration with the rates being 41% and 74% for TBI patients in general, and 27% and 32% for post-concussion patients (Engberg & Teasdale, 2004; Fox, Lees-Haley, Earnest, & Dolezal-Wood, 1995).

1 The US constitute only around 5% of world population but these numbers cannot be extrapolated because the offerings and cost of offerings vary a lot between regions.

1.1.3 Rehabilitation

Relatively effective[2] interventions have been developed for the more specific impairments. For example, more than 18 intervention techniques have been proposed for the treatment of visuo-spatial neglect (Luaute, Halligan, Rode, Rossetti, & Boisson, 2006). One widely used and effective intervention is prism adaptation (Rossetti et al., 1998) where patients wear goggles that shift the visual field to the right, thereby increasing awareness of the world to the left of the centerline in the visual field. Other techniques include compensatory behavior such as trunk rotation, visual scanning training. Limb motor dysfunction can be effectively improved using constraint induced movement therapy in which patients are prevented from compensating with their functional extremities (Taub, Uswatte, Pidikiti, & others, 1999). Aphasia/dysphasia can be ameliorated with techniques from speech therapy, e.g. constraint induced language therapy (Pulvermüller et al., 2001).

These are specific remedies for relatively specific deficits. High-level cognitive impairments are less tangible, making it harder to notice mild impairments (e.g. to distinguish it from pre-injury low intellect) but also harder to rehabilitate severe impairments (Park & Ingles, 2001).

1.1.4 Cost-effectiveness of rehabilitation of high-level cognition

Several interventions have been proposed to improve impaired high-level cognitive sequelae of brain injury. In this section I will informally review some non-computerized interventions and evaluate their cost-effectiveness. For each intervention type, all relevant papers on the first 5 Google scholar pages and their references were scanned, totaling to approximately 100 scanned abstracts. Only studies with adult brain injured subject, parallel control groups, and accessible full text were included. All performance-based cognitive outcome measures (questionnaires not included), that were hypothesized to be improved from the intervention, were considered. From these, the average standardized mean difference (SMD) were calculated, standardized using the average of the pretest standard deviations. Statistical noticeability ($p < 5\%$) was scored for group □ time interaction terms only.

Interventions include (1) Attention Process Training, a face-to-face intervention with a

2 Here “effective” means that less than 40 hours of intervention is needed to produce large effects (SMD > 0.8).

number of therapist-administered exercises and strategies to improve attention, (2) Mindfulness Based Stress Reduction which claims to improve attention per se, (3) physical exercise, and (4) pharmaceuticals. The latter two target cognition in general whereas the first two target attention specifically. For the sake of brevity, I take the results from published effect sizes at face value without subjecting them to theoretical and methodological scrutiny.

Since resources are limited in healthcare, the best intervention is the one which yields large effects using the fewest resources. In clinical research, the primary resource drain is therapist hours so that may be used as an index of the cost of an intervention. The Number Needed to Treat (NNT) statistic can be calculated from SMD (Furukawa & Leucht, 2011) and may be used as an index of the effectiveness. Combining these to an index of cost-effectiveness, I calculate an Hours Needed to Treat (HNT) statistic by multiplying the Number Needed to Treat (NNT) statistic with the number of therapist hours used in each study (HNT = NNT x therapist-hours).

None of the four intervention types have satisfactory cost-effectiveness with respect to high-level cognition (see Table 1). It would take an expected 586 therapist hours for Attention Process Training to make one patient show a more favorable result than would be expected in the control group. This translates into 73 full 8-hour workdays, planning and administration not included. 319 hours are expected for Mindfulness Based Stress Reduction and 930 hours for physical exercise. The average SMD for pharmaceuticals (amantadine, methylphenidate, and sertraline) was 0.175 which translates into a NNT=19. Pharmaceuticals take little therapist time but one positive outcome out of 19 patients is not satisfactory when considering medicine costs and side effects for the remaining 18 patients.

I conclude that Attention Process Training, Mindfulness Based Stress Reduction, exercise, and pharmaceuticals are not cost-effective for cognitive improvement following acquired brain injury (see also Cumming, 2013). This is not always apparent from published research and reviews which treat a single uncorrected statistical noticeable outcome ($p < 5\%$) as synonymous with effectiveness even though p depends just as much on sample size and sampling intentions (e.g. number of outcome measures and stopping rules; see Kruschke, 2011; Nickerson, 2000).

There is a clinical need for treatments which require fewer resources, i.e. few therapist hours for behavioral interventions, and with larger effect sizes. This is the state of affairs that this thesis aims to improve. For comparison, see Table 13 in the discussion for similar statistics on the two intervention papers in this review.

Publication	Treatment	p < 5%	Hours	SMD_c	NNT	HNT
Barker-Collo (2009)	APT	6 of 10	13.5	0.33	9.57	129.2
Park (1999)	APT	4 of 6	40	0.13	26.08	1043.2
Johanssen (2012)	MBSR	3 of 5	64	0.33	9.57	612.5
Azulay (2013)	MBSR	3 of 4	20	0.22c	14.9	298.0
McMillan (2002)	MBSR	4 of 12	3.75[c]	0.09	38.26	48.7
Studenski (2005)	Exercise	0 of 2	54	0.11	31.1	1679.4
Ploughman (2008)	Exercise	0 of 5	2.8	0.01	355.7	996.0
Quaney (2009)	Exercise	1 of 5	18	0.47	6.45	116.1
Meythaler (2002)	Amantadine	0 of 4	%	0.67	4.3	%
Schneider (1999)	Amantadine	0 of 4	%	0	∞	%
Lee (2005)	Methyl-phenidate	1 of 10	%	0.23	14.2	%
Lee (2005)	Sertraline	1 of 10	%	-0.20	%	%

Table 1: *Effect sizes, consistency of outcomes and economic feasibility of different interventions for cognitive rehabilitation. MBSR: Mindfulness Based Stress Reduction. APT: Attention Process Training. SMD_c: the controlled standardized mean effect, $SMD_c = (\Delta_{treat} - \Delta_{control})/((\sigma_{treat} + \sigma_{control})/2)$ where σ is the standard deviation of pretest scores (Cohen, 1992). NNT: Number Needed to Treat calculated from SMD_c, assuming a controlled event rate of 20% (Furukawa & Leucht, 2011), HNT: Hours Needed to Treat (HNT = NNT x hours). [a][positive/total] outcome measures. [c]This study had a total of 2*10 hours of meditation when including home-exercises, corresponding to an HNT of 715.5.*

1.2 Inferring the aptitude of cognitive constructs from behavioral tests

The vast majority of cognitive research uses standardized tests to make inferences on cognitive constructs. In clinical research in particular, it is common to assign cognitive labels to tests. For example, the Stroop task is often used to measure "inhibition", PASAT to measure "attention", digit span to measure "working memory", and the Wisconsin Card Sorting Test to measure "executive function".

This thesis is no different. In this section, I will discuss what can be concluded from such inferences and the assumptions it entails. I will argue that it is a logical error to assign a single psychological construct to each particular neuropsychological test and that it is also too strong a simplification to be justifiable. Rather, constructs are useful systematizations of probabilistic relationships between behavioral performance on a set of neuropsychological tests. We need not commit ourselves to the existence or non-existence of such constructs, nor to the idea that they can be truly measured.

1.2.1 Psychology as ill-posed reverse engineering

Psychology can be seen as reverse inference or reverse engineering[3] of the mind from behavior (Dennett, 1994; Pinker, 1997; Poldrack, 2006). In general, reverse engineering is the act of reconstructing an unknown mechanism by observing its input and output. The mechanism is a "black box" with interiors that will remain forever unobserved. In psychology, the input is stimulation, the mechanism is cognition, and the output is behavior. In the broadest sense, behavior includes movements, blushing of the skin, electromyographic (EMG) activations, electroencephalography (EEG), brain metabolism (usually measured using BOLD), and other overt reactions to input (see Henson, 2005, for an argument for this definition of behavior).

3 Forward inference is knowing the generating mechanism and inferring the output it would generate given a specific input. Reverse (or "backward") inference is knowing the input and the output and inferring the mechanism that mediated between the two. Determinism is usually assumed in both forward and reverse inference.

A neuropsychological test is just this kind of inference. In digit span, for example, we read a series of digits to the subject (input) and observe how many digits were recalled correctly (output). Since digit span tends to follow a (probabilistic) Guttman scale[4], we may use these data to infer that this reduction in output relative to input was caused by a capacity limit in the cognition of this subject.

Contrary to the well-posed problem of forward inference, reverse inference is often ill-posed in that there is no unique solution. Many different mechanisms could generate the same observed input-output relations[5]. To give just one example from cognitive psychology: capacity limits on short-term storage of information has been argued to be 7 items (Miller, 1956), 4 chunks (Cowan, 2000), 4 objects (Luck & Vogel, 1997), 2 seconds for sound (Baddeley & Hitch, 1974), 1 chunk (McElree, 2001), and many more (see Cowan, 2005, pp. 139-164 for a list). Numerous theories have sought to describe the mechanism which has the property of such capacity limits (e.g. Baddeley & Hitch, 1974; Norman & Shallice, 1986; Cowan, 1988). These theories disagree on the cause of the capacity limit (decay, interference, etc) and the unit of the capacity limit (items, chunks, objects, time, etc.). It is clear that it is difficult to raise any one of these theories to the status of "absolute truth which solved the reverse inference problem" with respect to capacity limits on memory. This problem is particularly strong for high-level cognition which, by definition, is involved in all behavior in flexible ways and thus is difficult to isolate using input-output relations alone. Lower-level computational processes, such as detection thresholds for visual objects or auditory sounds, allows for tighter inference.

To conclude: cognitive constructs such as attention and working memory are unobservable. As a consequence, they are not uniquely defined epistemologically speaking. This has implications for how we approximate the aptitude of such constructs.

1.2.2 Cognitive constructs group correlated behaviors

It is circular reasoning to say that the performance on a given behavioral test directly channels a cognitive construct (e.g. "WCST measures executive function"), since that construct was inferred from behavior in the first place[6]. This logical fallacy can be avoided

4 A Guttman scale has a sharp cutoff between successes and failures, e.g. 1, 1, 1, 1, 0, 0, 0. This can be made probabilistic as in e.g. Rasch analyses where sequences like 1, 1, 1, 1, 0, 1, 0, 0, 0, 0 can be interpreted as a sharp cutoff with random noise.

5 This is known as the problem of multiple realizability in philosophy of mind.

6 This circular reasoning is especially apparent in psychiatric diagnoses where it is often said that the diagnosis is the cause of behavior. E.g. "he is sad because he is depressed". In the DSM and ICD systems, the diagnosis is based on behavioral symptoms, e.g. sadness for a certain period.

by turning the statement into a probabilistic one about relationships between behaviors, e.g. "Performance on WCST correlates with performance on other tests of executive function". Here, the cognitive construct is merely an "epiphenomenal label" which identifies correlations between behaviors that fall under this label. To give a more explicit example, I propose the following translation:

- Cognitive language: He is doing poorly on a working memory task. His working memory is probably impaired.
- Translates to: He is doing poorly on a task which is a member of the category "working memory". Therefore he will probably do poorly on other tasks and behaviors in the category "working memory" since tasks covary within categories.

In other words, cognitive constructs is a systematization of behavior. The above translation puts dichotomous reasoning in the background and brings probabilistic reasoning in the foreground by substituting an all-or-none categorical construct ("working memory") with an (imperfect) correlation.

This is not new. It follows directly from the way that we assess the construct validity of neuropsychological tests using structural equation modeling (most frequently factor analysis). Here, individual tests are clustered together using mathematical procedures that ensures maximal shared variance within clusters and minimum shared variance between clusters. Such clusters are not uniquely defined. Therefore it is typical to consider several probable cluster configurations in these publications. Furthermore, the labeling of the factors is an entirely subjective act based on the compound face validity of the items in each derived factor.

1.2.3 Any test represents multiple constructs

The above highlights the non-unique relationship between test and construct. If a test reflected a single construct, then performance on two tests of different constructs should be uncorrelated. This is seldomly the case[7]. For example, the Wechsler Adult Intelligence Scale has a common intelligence factor which explains a large proportion of the variance in individual tests and derived indices (Working Memory, Perceptual Reasoning, and Verbal Comprehension; see e.g. Benson, 2010; Holdnack, 2011; Canivez, 2010). Even though these tests are often labelled according to the derived indices under which they are

That turns the sentence into "he is sad because he is sad". Likewise, the WCST utterance can be turned into "WCST measures WCST performance".

7 There are several papers who do make claims about conditional independence between factors, by factoring out variance from other constructs, in which case one has already assumed more than one construct as explanatory variables for the behavior. Conditional independence is not an argument against what I propose here.

subsumed, this shared variance show that they are not independent. Given that the correlation to the common intelligence factor is not 1, the outcome on each test is determined by multiple constructs. Similarly, Roca et al. (2010) found that fluid intelligence accounted for all differences between patients with frontal lesions and matched healthy controls, indicating that all tasks depended to some degree on fluid intelligence in addition to their more local domains. The same has been found in numerous other latent variable analyses on other cognitive outcome measures (e.g. Engle et al., 1999; Conway et al., 2002; Kane et al., 2005; Friedman et al., 2006). All of these findings are at odds with a 1-to-1 test-construct relationship.

The multiplicity of constructs involved in performance of a particular test can also include more task-specific factors. Consider, for example, the digit span test in which subjects are asked to repeat gradually longer sequences of digits. For auditory input, the number of recalled digits would be reduced by impressive aphasia and impaired hearing. For visual input, blindness and neglect would reduce the output. For both modalities, dyscalculia, memory capacity, attention, and expressive aphasia (for verbal output) would limit output, to name a few. Therefore digit span performance could be bottlenecked by any of these or any combination of these - not just a memory capacity limit[8].

Again, I propose a translation from categorical language to probabilistic reasoning:

- Cognitive language: Digit span is a working memory task.
- Translates to: digit span shares more variance with tests labeled “working memory” than tests which are not labeled “working memory”.

1.2.4 Choosing an explanation among ambiguous constructs

So far I have argued that there are many constructs which could be said to mediate between stimulation (input) and performance (output) on a single test and that many different constellations of such constructs could give rise to observed behavior. It is an ill-posed problem to objectively choose among such constellations. Where we reach the limit of inference we can, however, use a normative criterion to choose between competing models. To this end, Morgan’s canon is often used as a guiding principle for comparative psychology (see e.g. Hampton, 2009):

8 The digit span test is used to measure attention, working memory, or short term memory, depending on publication. This speaks to the ambiguity of one-to-one relationships between tests and cognitive constructs.

> *In no case is an animal activity to be interpreted in terms of higher psychological processes if it can be fairly interpreted in terms of processes which stand lower in the scale of psychological evolution and development. (Morgan, 1894, p. 53)*

Morgan's law can be said to enforce a law of parsimony in terms of the predictive space of the proposed explanation[9]. Since concepts lower on the psychological hierarchy tend to process a smaller set of stimuli in a more rudimentary way, the predictions from low-level models are more restricted. This makes low-level models more exclusive and therefore more parsimonious. Concepts higher in the psychological hierarchy tend to be more versatile and have a larger prediction space, thus being more complex (Lee & Wagenmakers, 2014).

1.2.5 Summary

In short, we may use "cognitive language" for convenience and the translations for scientific accuracy. The translations shift the focus away from a language which implies that cognitive constructs exist as known entities with known aptitudes, towards a science based on shared variance between tasks with cognitive labels, thus preventing false assumptions and circular reasoning. There is nothing lost in this translation and as such it should be an uncontroversial reminder. It still allows for reasoning about hitherto unobserved behaviors (theoretical predictions) and the "underlying" effectiveness of interventions given observed behavior.

9 It is captured in the idiom "A model which predicts everything predicts nothing" (Lee & Wagenmakers, 2014). This is a different approach than the more well-known model selection in maximum-likelihood statistics which use the number of parameters as a measure of complexity (for models where covariates share the same residual, the number of parameters is actually proportional to the predictive complexity), using e.g. AIC og BIC.

1.3 Probabilistic inference: Bayesian inference does what we want to do. Null Hypothesis Significance Testing does not.

Empirical science use samples to draw inferences about the parameters of a population, i.e. the parameters which generated these samples. In the context of rehabilitation research, these parameters are often effect sizes of interventions. Probabilistic[10] inference is the act of narrowing down the values of parameters in a mathematically sound way given the data and a model of the relationship between parameters.

Since probabilistic inference side with the experimental method as the most important tool to draw conclusions about the efficacy of various rehabilitation methods, and since I will deviate from mainstream Null Hypothesis Significance Testing practices, it deserves some attention here. But since this thesis *uses* probabilistic inference but is not *about* probabilistic inference, I will avoid presenting a full mathematical and philosophical analysis of methods for probabilistic inference although it is a subject near and dear to my heart. I have tried to keep this section relatively short and conceptual in order to motivate the choices of inferential procedures without losing focus from the real task at hand. The footnotes present expanded justifications for the rather superficial claims in the main text.

1.3.1 p-values and confidence intervals have low inferential value

p-values do not reflect the strength of evidence against the null hypothesis nor the inverse of the strength of evidence for the alternative hypothesis (Nickerson, 2000; Goodman & Royall, 1998; Cohen, 1994).

The fathers of current statistical practices acknowledged this fact. Ronald Fischer proposed p-values as an "index" which approximately co-varied with the real probability (Fisher,

10 "Probabilistic" inference is contrasted to normal logical inference in that we assign numerical probabilities to parameter values instead of TRUE / FALSE labels. So given incomplete data we may infer probabilities but not truth. This is also the reason why data does not "prove" a hypothesis (read: assign a TRUE label) but rather "supports" them (read: increase the probability of it being true).

1959, p.44.)[11]. Neyman & Pearson developed a statistical framework which avoids assigning probabilities to hypotheses altogether because they saw probabilistic inference as inherently problematic. The Neyman-Peterson framework simply serves to control the rate of false positive claims and false negative claims in the limit of infinite perfect replication attempts (Neyman, 1957).

Fisher and Neyman-Pearson fiercely disagreed on how to use data to falsify or support models[12]. Yet, while the debate was ongoing, the Fisherian and Neyman-Pearsonian frameworks appeared in an inconsistent hybrid version[13]. This "bastard child" which

11 A nice selection of quotes from Fisher's book is collected in the contemporary book review by Pitman (1957)

12 Fisher was advocating "inductive reasoning" using essentially a probabilistic version of Popper's falsification principle: quantify how likely the data is under the null and use this as your inference without directly accepting or rejecting any model. There is only probabilistic falsification; no support of an alternative hypothesis. The p-value should be interpreted as an approximate measure of the strength of falsification. Any decisions can then be made based on these beliefs afterwards.

Neyman advocated "inductive behavior" based on the observation that evidence is used to assign TRUE/FALSE labels to discrete hypotheses in scientific practice. This kind of behavior will sometimes be wrong (type-1 and type-2 error) and controlling the rate of these error-behaviors will lead to proper inductions in the long run. I find that the discrepancies between Neyman's and Fisher's frameworks are concisely captured by Fisher (1956) and Neyman (1957) in which they launch fierce attacks on each other.

The key disagreements are these: Fisher had no acceptance of an alternative where Neyman-Pearson did. Fisher distinguished decision from inference where Neyman-Pearson did not. And Fisher regarded the p-value as a (approximate) quantitative measure of the strength of evidence (against the null) where Neyman-Pearson used a decision threshold (alpha).

I am most sympathetic to Fisher's arguments in this debate although I of course take distance to both frameworks for the reasons mentioned in the main text. My primary concern is that even if it may be common practice to use a measure of probability label hypotheses as TRUE or FALSE, induction from data is not logical but probabilistic and the dichotomous labels TRUE/FALSE are therefore unwarranted. The best we can do is to assign probabilities to the two (or redistribute it if they had prior probabilities). As a remark, I side with Fisher on the impossibility and strangeness of relying on infinite hypothetical perfect replications for inference or as Jeffrey's put it: "*A hypothesis that may be true may be rejected because it has not predicted . . . results which have not occurred."* (cited in Goodman & Royall, 1988, p. 1569)

13 Most papers on the NHST controversy identify contemporary textbooks as the likely source of the hybrid version. However, this is not supported by historical investigations of textbooks from the period (Halpin & Stam, 2006; Huberty, 1993) which point to small-sample research papers as a possible source.

neither Fisher nor Neyman would have accepted[14] is now widely known as the Null Hypothesis Significance Testing (NHST, see Nickerson, 2000, for an introduction). NHST is now applied as if it was a sound probabilistic inference procedure. In short, NHST consists of the following steps:

- Define a Null hypothesis (usually the Nil hypothesis of no difference in population means)
- Decide on a threshold (alpha) for the accepted rate of false positives claims.
- Collect data.
- Calculate the likelihood $p = P(data \mid H0)$ that the current data were generated under the null hypothesis.
- If p < alpha, claim that the hypothesis is supported. If p > alpha, claim nothing.

A major practical problem with NHST is that the null hypothesis may be more likely than the alternative, even though p < alpha (Berger & Sellke, 1987; Rouder, Speckman, Sun, Morey, & Iverson, 2009). In fact, it has been shown that if the H0 and H1 are equally likely a priori, a p-value of 5% corresponds to a 29 % probability of H0 being true within the logic of the NHST-framework (Sellke, Bayarri, & Berger, 2001) and between 33% and 82% (depending on the sample size) using a Bayesian analysis as reference (Berger & Sellke, 1987).

Several philosophical problems add to this, including (1) that p=alpha+0.0001 and p=alpha-0.0001 are clearly not qualitatively different but we are forced to make that statement in NHST. (2) The strangeness of doing inference on perfect replications which has never been done and probably never will. (3) the impossibility of such perfect replications (see e.g. Fisher, 1956) and (4) that differences in an experimenter's prior sampling intentions lead to different p-values for the same dataset independently of the data (Goodman & Royall, 1988; Kruschke, 2011). Many of these issues were raised in the Fisher versus Neyman-Pearson debate in the 50's but they remain largely ignored in the scientific empirical literature today which ritually follow the NHST procedure.

14 For example, both Neyman-Pearson and Fisher thought that the null hypothesis should be well qualified and thus seldom would be 0 (what we now call the Nil hypothesis). Once the null is defined, the strength of evidence required to accept or reject some hypothesis should also be decided on a per-study basis. In Neyman-Pearson this plays out as alpha and beta values which essentially constitute a loss function on type-1 and type-2 errors in this particular context. Fisher thought that different levels of p would be required to reject the null, depending on the context. Neyman-Pearson would accept the null whereas Fisher would not accept anything.

Today NHST almost always test the Nil hypothesis (mean=0) against a universal p<5% threshold (corrected or uncorrected) which is used to accept the alternative but not the null. This is clearly at odds with both Fisher and Neyman-Pearson.

Confidence intervals has been suggested as an improvement over p-values (Gardener & Altman, 1986; Cumming & Finch, 2005). However, confidence intervals are simply re-expression of p-values centered on a mean and scaled by the observed dispersion (Kruschke, 2011; Feinstein, 1998). As a consequence, they inherit the same problems as p-values. In fact confidence intervals are often used in the same way as p-values to judge effect sizes to be significant or not-significant (Coulson, Healey, Fidler, & Cumming, 2010; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2014) and used in this way they add nothing new. I will return to the definition of confidence intervals in a later section**.**

For the reasons just presented, I conclude that p-values and confidence intervals have low value as a measure of evidence for different hypotheses and effect sizes. In the next section I will argue that Bayes Factors (BF) and Highest Posterior Density intervals (HPD) have much higher value for these inferential goals. If used right, Maximum-Likelihood based Likelihood Ratio Tests (LRT) may approximate a Bayesian model comparison.

1.3.2 Bayesian inference has high inferential value

Bayesian inference[15] is the branch of mathematical probability theory which is based on the Kolmogorov Axioms[16]. Bayesian inference deals with conditional probabilities between parameter values. Two parameters are conditional if a change in one of them changes our knowledge about the other, however little that change might be. In empirical science, we often condition parameter values on data: conditioning on data changes our belief in what parameter values might have generated these data. But c.f. the discussion about how to "measure" cognitive functions, there are many underlying parameter values which could have generated a given dataset. The bayesian solution to this ambiguity is to assign

15 Bayesian inference is named after Thomas Bayes (1701-1761) although it probably owes more credit to Pierre-Simon Laplace (1749-1827) and Andrey Kolmogorov (1903-1987).

16 These axioms were originally presented as 5 axioms by Kolmogorov's "Foundations of the theory of probability" (1956, p.2) but has since been reformulated to three axioms (see e.g. Russel & Norvig, pp. 484-489 for a very clear formulation):

- 1. The probability of a state is a non-negative real number.
  $P(state) \in \mathbb{R}, P(state) \geqq 0$
- 2. The probability that one out of all possible states occur is 1.
  $SUM(P(state)) = 1$
- 3. The probability that one out of several disjoint states occur is the sum of the probabilities of the respective states.
  $P(state1 \mid state2) = P(state1) + P(state2)$

From axioms 1 and 2 it follows that $0 \leqq P(state) \leqq 1$. Note that in most texts the word "event" is used where I write "state". I prefer the latter since the former has connotations of real-life events, i.e. something local at a specific time.

probabilities to each value that the underlying parameter(s) might have[17]. This allows us to reason about such "hidden" values to the best of our knowledge about them. In this way, bayesian inference is well suited to get the most out of the reverse inference problem.

Continuous parameters can have an infinite number of different values, so to simplify how we describe our current knowledge state, we use distributions. For example, the normal distribution simply takes two parameters: the mean and the variance. Likewise, the uniform distribution takes just two parameters: the minimum and the maximum.

1.3.3 Three prior distributions for effect sizes

One of the most disputed features of Bayesian inference is this necessity of prior information for the calculation of conditional probabilities. Although it arises as a pure mathematical necessity, many intuition-appealing interpretations of the math has been put forward (see e.g. Edwards, 1963). One of the most mathematically true arguments is this: you can only reason about that which is conceivable (i.e. has an a priori non-zero probability). If it is inconceivable (has a prior zero probability), no amount of data could make it conceivable. So everything, parameters and parameter values included, has to have a non-zero probability before you can even begin to reason about it and we call this the prior probability.

Again, we often use distributions to specify our prior knowledge. We usually distinguish between narrow or "informative" priors and wide or "uninformative" priors. Informative priors concentrate probability on a relatively small range of states, deeming some states much more likely than others a priori. Uninformative priors distribute probability on a relatively large proportion of states, deeming them approximately equally likely.

I will use three different priors in this thesis. They are illustrated in Figure 1. For parameter estimation, I use an informative Normal(mean=0, sd=0.5) prior on SMD which puts 95% probability on the interval -0.98 SMD to 0.98 SMD. This is informed by the observation that effect sizes in behavioral intervention research tends to be below 1 SMD.

For model selection, I rely on two other priors, primarily because they have nice computational properties and because they have been widely used in the scientific literature. The JZS prior has been proposed as a "dealt bayes factor" for model selection of linear models (Rouder & Morey, 2012). It is a Cauchy(0, sqrt(2)/2) which puts 95%

17 I have often seen it said that the difference between frequentists and bayesians is that frequentists think that there is just one real parameter value while bayesians think that it is distributed. This is not accurate since bayesians also assume (per the Kolmogorov axioms) that there is one real parameter value. But the bayesian belief in what that value is distributed given incomplete knowledge. Frequentists postulate a single value given data and then make inferences on the frequency with which they'd be wrong in making such statements.

probability between -8.9 and 8.9 SMD. This is clearly way outside what could reasonably be expected a priori in intervention research but due to the shape of the Cauchy (which is just a t-distribution with 1 degree of freedom), it still has a great deal of mass around zero with 60% in the interval -0.97 to 0.97 SMD (see Figure 1).

Lastly, I use the very wide unit information prior where the other bayes factors are not possible to compute (see a later section). The unit information prior can be calculated based on the Bayesian Information Criterion of two models estimated using the Maximum Likelihood method (Wagenmakers, 2007). It is a normally distributed prior that corresponds to the information of having a flat prior and one observation at the maximum likelihood estimate. The unit information prior depends on the data[18] so it cannot be illustrated correctly. To give an impression, I have plotted a wide Normal(0, 5) prior in Figure 1. That prior has approximately the same 95% credible interval as the JZS prior, but puts much more probability on extreme effect sizes.

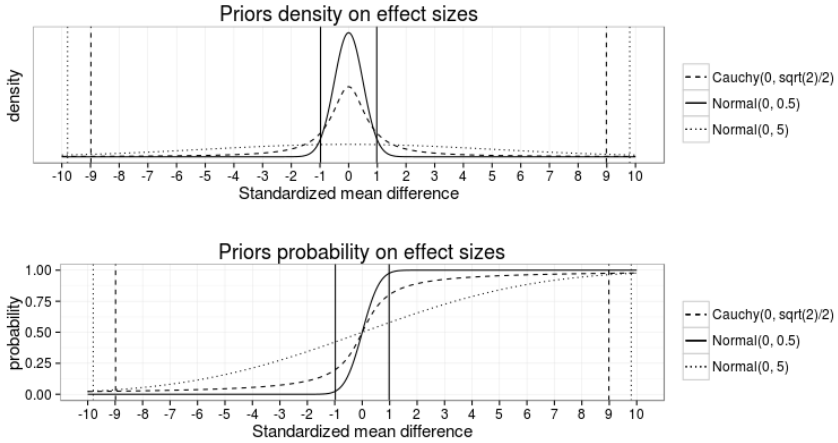

Figure 1: *Density and probabilities of different priors as a function of effect size. Vertical lines mark the 95% credible interval for each prior. The vide Normal(0, 5) prior is illustrated to show how the Unit Information Prior might behave compared to the other two priors.*

18 And is therefore not proper

A part of the worry about the need for priors relative to the NHST ritual is that there is no correct prior. In principle, this makes perfect scientific consensus impossible since each scientist has his/her own prior knowledge[19,20]. However, a rule of thumb for scientific reporting is to choose a prior which is acceptable to a skeptical audience. That will usually be an uninformative prior or one that is conservative with respect to the expected outcome.

1.3.4 The posterior distribution as the parameter estimate

When conditioning the model on data, the resulting distribution of probabilities across parameter values is called the posterior distribution. The posterior distribution is a complete description of the knowledge we have about what parameter values might have generated the data, i.e. our inference on the parameters. Often, the posterior distribution follows a well-known distribution and can simply be described using that distributions' parameters, i.e. the mean and standard deviation for a normal distribution.

19 This, by the way, is the point of departure for the epistemology of Chalmers (2011) in which rational subjects have a justified true belief about a proposition if that proposition follows from his/her experience. Chalmers calls this prior endorsement. It is a challenge to classical analytical philosophy in which truth can be fully uncovered and agreed upon by rational subjects irrespective of their prior experiences. I am sympathetic to the view of Chalmers here.

20 This is the reason that Bayesian inference has been called "subjective" and is the major source of criticism from the frequentist stance which claims to be "objective". This criticism is largely unfounded, though. Note that both Fisher argued that p-values should be interpreted qualitatively with respect to the phenomenon at hand and Neyman-Pearson argued that alpha-levels should be set per-experiment. I.e. the founders of modern what became NHST strongly endorsed the necessity of prior information in inference. In that sense, Bayesian inference is less subjective since the use of prior information is principled and mathematical. This is the philosophical argument. The prior is also an effective means to fix the problems with the NHST neglection of base rates which leads to many many erroneous inferences since rare events is conflated with evidence against the null. The most famous of which is the problem of mammography and breast cancer which was introduced by Eddy (1982) and subjected to thorough analysis by Gigerenzer & Hoffrage (1995). Cohen (1994) is maybe the most famous single published rebuttal of the failure to take base rates into account in NHST.

However, we are often interested in knowing the interval in which we can expect the parameter to be in. The credible interval is the interval with the highest posterior density. As such, the credible interval is free from any assumptions about the distribution of the posterior. Many people interpret the confidence interval as if it was a credible interval (Coulson, 2010), but consider the difference in what the two intervals are:

- **The 95 % Confidence interval** is the interval which, if calculated in infinite perfect replications of the experiment, would contain the true parameter value in 95 % of these calculations.
- **The 95 % Credible Interval** contains the true parameter value with 95 % probability.

1.3.5 Model comparison using Bayes factors

Since a hypothesis is most frequently deduced from competing theories, multiple statistical models are at play. Testing a hypothesis comes down to evaluating which theory-based statistical model is most likely given data from the experiment. Procedures to this end are called model selection or model comparison[21].

I will be using Bayes factors for model comparison. A Bayes factor is the odds ratio between the posterior probability of the two models in question. It is explained in more detail in the three papers but briefly, a bayes factor of 5 in favor of model #1 means that it is 5 times more likely to have generated the observed data than model #2 which conversely is then 5 times less likely to have generated the observed data than model #1. It does not matter if we consider one of them a null model - we can quantify evidence for all models.

Bayes factors can be computed using the product space search method (Lodewyckx et al., 2011; Han & Carlin, 2001), where all models of interest are included in a grand model, connected by an indicator parameter. When conditioning on the data, the posterior distribution over the indices of the indicator parameter is the posterior probability of the models. Conceptually, this simply means that instead of having a fixed 100% prior probability that the parameters exist, we now make inference on that probability.

Others advocate the use of posterior predictive checks for model comparison instead of bayes factors (Gelman et al., 2013; Gelman & Shalizi, 2012). This debate is beyond the scope of this thesis (see Morey, Romeijn, & Rouder, 2012, for a reply). Bayes factors do

21 "Selection" implies that a binary TRUE/FALSE label is assigned to each model as is common in step-wise frequentist model selection procedures where parameters are included/excluded using p-values with respect to an alpha threshold. "Selection" is a misleading word in a bayesian context since the data is used to assign probabilities to each model rather than accepting or excluding them. "Model comparison" is a more appropriate term.

have the advantage that they are axiomatic while posterior predictive checks are not. In either case, I have used both for all the statistical models in this thesis but for the sake of brevity, I only report the bayes factor.

In practice, however, the product space method can be computationally expensive to the extent that it is not feasible. This is also true for some of the models in this thesis, e.g. in the meta-analysis model and when testing multiple parameters simultaneously in the hypnosis model (See Appendix B). For simple models, the Savage-Dickey method can be used to estimate the Bayes factor from the posterior distribution of the largest of two nested models (Wagenmakers et al., 2010). Since it is inappropriate for the more complicated models used in this thesis, it is not reported here. A Bayes factor estimate can also be computed from the difference between the Bayesian Information Criterion (BIC) of the models in question, as discussed in the next section.

1.3.6 The Likelihood Ratio Test: a Maximum-Likelihood Bayes factor

There is a Maximum Likelihood equivalent to the Bayes factors analysis: the Likelihood Ratio Test (LRT). The likelihood is the probability of observing data under a given model, $P(D \mid M_i)$, i.e. how well the model accounts for data. The likelihood ratio is simply the ratio between the log-likelihoods of two models, e.g. a full model and a null model. Since the likelihood of the data given a model is proportional to the model evidence given data, likelihood ratios are proportional to the bayes factor. However, the likelihood ratio is based on a point estimate while the bayes factor integrates over the full parameter space. In that way, the bayes factor accounts for model complexity while the likelihood ratio is overfitted. However, a Bayesian Information Criterion can be computed by combining the likelihood ratio and a penalty for the number of parameters in the model and that can be used to calculate a bayes factor, $BF_{BIC} = \exp(BIC_{full} - BIC_{null})$ (Wagenmakers, 2007). This bayes factor assumes the very uninformative unit information prior as described above.

Today's academic publication landscape demands p-values so the p-value of the LRT is reported together with the bayes factor in the papers. I have presented reasons to ignore that p-value.

1.3.7 Choosing Bayes: what are the implications for this thesis?

Let us assume that if I had used NHST I would commit all the common errors of NHST, i.e. use maximum-likelihood parameter estimates, interpret the 95% confidence interval as the interval within which the parameter is with 95% probability, and think that $p < 5\%$ is evidence that the alternative is true and the null is false.

Since I analyse linear models on approximately normal data, there is a large correspondence between NHST and Bayesian estimates. The following differences should be expected:

- Bayes-derived SMD is smaller than Maximum Likelihood derived SMD. The SMD posterior is not normally distributed since the posterior of the population SD has a positive tail. Therefore the SMD will have more mass on smaller effect sizes than an ML estimate which assumes that the SD is normally distributed.
- BF-based quantification of evidence against the null is more conservative than p-based ditto (Wetzels et al., 2011), so I will be less likely to embrace full models than had I used NHST.
- I will quantify evidence for the null. This is not possible in NHST.

2 Computer-based cognitive rehabilitation

2.1 N-back Prologue

I will keep this introduction short since the paper is relatively self-contained. The conclusions advanced in the paper will be subjected to scrutiny in the final epilogue after the review paper.

The N-back study was conceived in 2011 inspired by a series of papers by Susanne Jaeggi and colleagues (Jaeggi et al., 2008, 2010, 2011) who demonstrated that training a single- or dual- N-back task caused transfer to untrained outcome measures in healthy subjects. At that time, I had not realized that the 2011-paper is a null finding masked as an effect and that the 2008-paper is deeply flawed (Shipstead et al., 2010; Redick et al., 2012). The study was initially simply designed to generalize these results to brain injury rehabilitation.

Within a few months during the summer of 2012 two research articles, two reviews, and a meta-analysis were published on the topic of computer-based training in general and the Jaeggi studies in particular. They uniformly discredited the idea that computer-based training have been demonstrated to cause improvements in high-level cognitive constructs such as fluid intelligence and working memory. Redick et al. (2012) published a thorough study on N-back training. First and foremost, Redick et al. uncovered that Jaeggi et al. (2008) had lumped together several very small experiments using different outcome measures and reported them as if they were the same experiment and the same outcome measure. Redick et al. then sought to replicate the findings with several methodological extensions. They stratified 140 subjects into three groups and tested them on 17 different tests at three timepoints. 75 subjects completed (43 % dropout rate). Although subjects improved on the training tasks, there was no differential improvement on the outcome measures at either timepoint, thus simultaneously discrediting the transfer-claim and the dosage-response claim put forward by the Jaeggi papers. Chooi & Thompson (2012) independently published a high-quality N-back study with 130 subjects starting and 93 completing (28 % dropout rate). Again, there was no differential improvement on the

outcome measures between the groups, thus discrediting the claim that N-back training causes transfer to fluid intelligence or working memory. Two reviews on computer-based working memory training (Shipstead, Redick, & Engle., 2012) and the CogMed software (Shipstead, Hicks, & Engle, 2012) was published simultaneously with the above studies. They convincingly showed that effects were too inconsistent between studies to establish a claim that they were effective. They were accompanied by a meta-analysis (Melby-Lervåg & Hulme, 2012) which found moderate effect sizes for near-transfer tasks but no transfer to untrained constructs (e.g. from "working memory training" to intelligence as Jaeggi and colleagues had claimed to show).

Still, other reviews concluded that the evidence speak in favor of computer-based "brain training" as an effective promoter of cognitive transfer effects (e.g. Morrison & Chien, 2010; Klingberg, 2010).

With these publications, I became aware that the research on healthy participants was seldom referenced in patient studies (and vice versa), meaning that the two research areas proceeded in parallel. For this reason, I decided to include a healthy group as well in order to build a bridge between research on brain injured and healthy subjects. I turned to bayesian statistics in the same period, so the prospect of a null finding was less scary since the null became interpretable. Certainly, support for the null hypothesis of no treatment effect is of very high clinical value since such a finding helps ruling out ineffective treatment methods. Finally, even the very slim chance of discovering positive transfer in a patient group would be overwhelmed by the benefit that it would pose for patients worldwide.

2.2 Training and transfer effects of N-back training for brain injured and healthy subjects.

Working Memory impairments are prevalent among brain injured patients. Computerized training targeting working memory has been researched extensively in non-injured populations but this field remain isolated from similar research in brain injured patients. We report the results of an actively controlled randomized controlled trial in which 18 patients and 19 healthy subjects completed training on the N-back task. The non-injured group had superior improvements on both training tasks (SMD=6.1 and 3.6) whereas the brain injured group improved much less (SMD=0.5 and 1.1). None of the groups demonstrated transfer to untrained tasks. We conclude that computerized training facilitates improvement of specific skills rather than high-level cognition in non-injured and brain injured humans alike. The acquisition of these specific skills seem to be impaired by brain injury. The most effective use of computer-based cognitive training may be to make the task resemble the targeted behavior(s) closely in order to exploit the stimulus-specificity of learning.

Randomized controlled studies on computer-based cognitive rehabilitation of brain injured patients go back to at least Sturm, Dahmen, Hartje & Wilmes (1983) and more than 50 RCTs have been published so far (see reviews by Lindeløv, submitted; Cha & Kim, 2013; Chen, Thomas, Glueckauf & Bracy, 1997). There is a much larger and growing literature on computerized cognitive training in non-injured individuals (see reviews by Shipstead, Redick & Engle, 2012; Melby-Lervåg & Hulme, 2012). However, these two fields have hitherto proceeded in parallel with no cross-talk or direct comparisons. Computerized cognitive neurorehabilitation could potentially expand its evidence base considerably if there are points of convergence to non-injured subjects. After all, all subjects are humans and all would benefit from improved information processing capacities.

Unfortunately, enhancement of domain-general cognitive functions such as working memory and attention has proven difficult in brain injured (Lindeløv, submitted, Park & Ingles, 2001) as well as non-injured subjects (Shipstead, Redick & Engle, 2012; Melby-Lervåg & Hulme, 2012) where the sum of evidence show a dominance of domain-specific effects of rehabilitation efforts, i.e. relatively little transfer to untrained material and contexts. There is an ongoing search for interventions which could promote far transfer, one of which is variations of computer-based cognitive rehabilitation.

There is a general distinction between training effects and transfer effects. This distinction is also known as domain-specific vs. domain-general and near-transfer vs. far-transfer. For example, working memory is said to be domain-general because it operates cross-modally on a wide range of stimuli domains and contexts (Baddeley, 2007; Cowan, 1988; Kanet et al., 2004). An improvement of working memory would then by definition lead to an improvement on all behaviors which rely on working memory. If that improvement was brought about by training a subset of these behaviors, such an effect be a transfer effect. In contrast, domain-specific processes apply to a narrow range of stimuli and contexts. The challenge for all studies is to provide convincing evidence that observed improvements are not mere training effects. For example, Westerberg et al. (2007) observed improvements on digit span and a spatial span tasks but both were a part of the training program so these results could be attributed to mere training effects.

In the present study we administered an N-back training procedure to a non-injured population and a patient population. The N-back has been used in numerous studies in non-injured participants with initial evidence of positive transfer (Jaeggi, Buschkueh, Jonides & Perrig, 2008; Jaeggi et al., 2010) but later replications have yielded null results (Jaeggi, Buschkuehl, Jonides & Shah, 2011; Redick, Shipstead & Harrison, 2012; Chooi & Thompson, 2012). An N-back-like intervention has only been administered to brain injured patients in a small study by Cicerone (2002) who found very large positive effects on untrained tasks in the order of a standardized mean difference of >2, although in a very different setting than the other N-back studies and with a sample size of only 4 patients in the intervention group. Studying the N-back task therefore serves the triple purpose of (1) vaguely replicating the Cicerone (2002) finding, (2) providing further evidence to the ambiguous results in the non-injured population and (3) to assess the extent to which evidence can be generalized across these two populations - at least for the N-back task.

2.2.1 Methods

Sample: We recruited 39 in-hospital patients with acquired brain injury (ABI) at Hammel Neurorehabilitation Centre and University Research Clinic. They did the training in addition to treatment as usual. 19 patients completed (see Figure 2). Additional inclusion criteria were (1) no aphasia, deafness, blindness or other disabilities which would prevent testing and training, (2) the patient should be able perform reasonably (d-prime > 1) at N-back level 1 and Visual Search level 2 at the time of recruitment and (3) and that the training did not interfere with the standard treatment as judged by the patient's primary therapists.

We recruited 39 non-injured participants who trained in their free time using their own personal computer. Twenty completed (see Figure 3). See Table 2 for descriptives on the recruited and final sample.

This study was approved by the local ethics committee, all subjects signed informed consent and participation was voluntary. Subjects were not reimbursed. Patients were informed that the training was designed to facilitate their cognitive recovery and non-injured subjects were informed that it was designed to boost their intelligence.

	Group	Included	Finished	N-back	Visual Search
Males / Females	ABI NI	31 / 8 18 / 21	13 / 4 8 / 10	6 / 2 3 / 6	7 / 2 5 / 4
Age in years	ABI NI	53.3 (10.4) 27.4 (10.3)	56.1 (6.3) 29.3 (11.3)	56.1 (5.6) 29.2 (11.1)	56.1 (7) 29.4 (11.9)
Days since injury	ABI	54 [28-94]	57 [33-95]	63 [35-89]	55 [33-95]
FIM cognitive[a] (0-35)	ABI	24 [22-28]	26 [23-28]		
FIM motor[a] (0-91)	ABI	82 [60-89]	82 [67-89]		

Table 2: *Sample descriptives at baseline for each group as frequencies or mean (SD). Included column is for all included subjects. Finished column describe subjects who eventually finished. N-back and Visual Search columns subdivides the finished columns to the two treatments. ABI=Acquired Brain Injury group. NI=Non-Injured group. FIM = Functional Independence Measure. [a]FIM scores at baseline were only available for 18 patients and should therefore be regarded as a rough indication of the patient group's functional level rather than a sample descriptive. It is not shown for individual treatments since there were only 4 of the patients in each group who finished that had a baseline FIM score.*

Design and randomization: This is a parallel-groups design with a 2 (patients/non-injured) x 2 (N-back/Visual Search) design resulting in four treatment arms. Participants entered the study continuously and were pseudo-randomly allocated to N-back and VS so that pre-test RAPM scores and ages were balanced. The studies on patients and non-injured subjects took place simultaneously but the stratification was independent for each group. Allocation of the first four participants was truly random. Participants were scheduled for a post-test when 20 training sessions were completed. Since more patients dropped out in the N-back group, more were allocated to this group in the continuous enrollment.

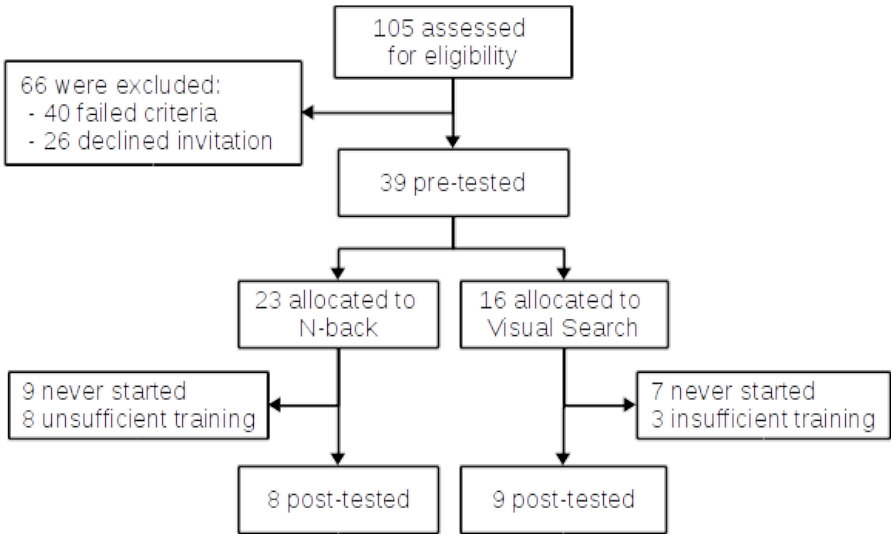

Figure 2: *Study design and flowchart for the patient group.*

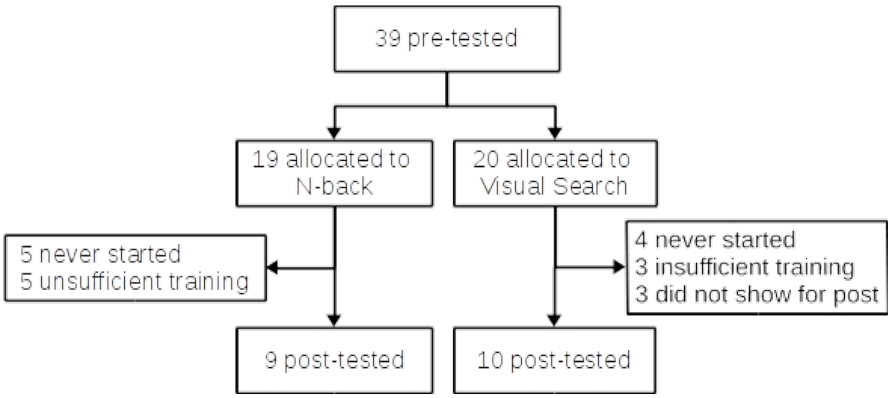

Figure 3: *Study design and flowchart for the non-injured group.*

N-back training: The N-back task consisted of a series of stimuli presented at 3 second intervals. The subject is instructed to press a key when the presented stimulus is identical to the stimulus N back in the sequence. There were 25 % targets per block and at most two consecutive targets. In order to prevent the formation of stimulus-specific strategies there were a total of 137 different stimuli as shown in Figure 4: 3 types of audio and 4 types of visual. A random selection of 8 stimuli from a randomly selected type was chosen for each block of training to ensure some interference between stimuli.

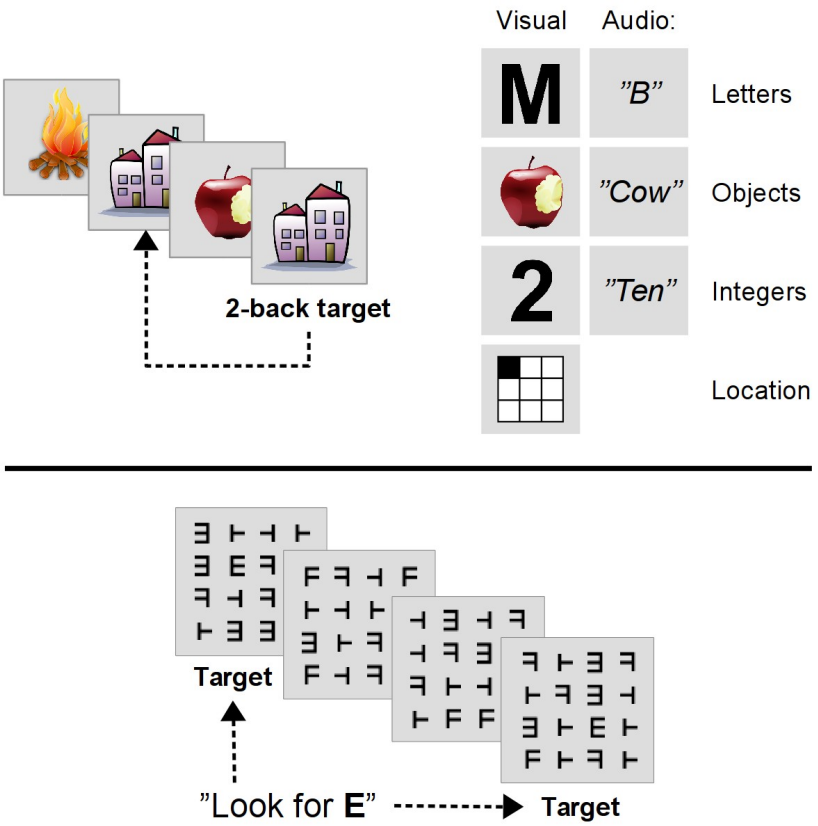

Figure 4: *The training tasks. Top left: 4 trials of the N-back task illustrated at 2-back level with image stimuli. Top right: seven different stimulus types. Bottom: 4 trials of the Visual Search task at level 4 (4 x 4 grid) and with "E" as a target. For each block any of the six different shapes would be picked randomly as target. Subjects were instructed to press a key on target trials.*

Visual Search training: Subjects were instructed to press a key if a target symbol was present in an N x N array of symbols. The target symbol changed from block to block but there were just 6 different symbols. The Visual Search (VS) task is unrelated to working memory (Kane, Poole, Tuholski & Engle, 2006) and served as an active control condition. It has served this purpose in other training studies (Redick, Shipstead & Harrison, 2012; Harrison et al., 2013). During training levels increased from N=1,2,3,4... etc. but we will use the number of items to be searched (N=1,4,9,16,...) as a measure of difficulty level throughout in this paper.

Both training tasks: Tasks were kept similar in all other respects to maximize the purity of the contrast. All participants trained 12 blocks of 20 + N trials on an adaptive N-back task for 20 days. Less than 10 blocks were considered an incomplete training day. Interstimulus interval was 3 seconds. Thus the full intervention consisted of 4.4 hours of constant training, breaks not included. Participants trained unsupervised in a laptop web browser.

Visual correct/incorrect feedback was given on response (hit/false alarm) and on misses in the end of every trial. To increase motivation, participants were awarded points after each block and given progressively more attractive titles. Feedback sounds were played on level upgrade or downgrade. In addition, participants would see a graph of their own progress.

Outcome measures: Subjects were tested on the following measures before and after training. Although each is related to many cognitive abilities, they are grouped according to the cognitive labels usually assigned to them:

- *Fluid intelligence:* Equal and unequal items from the Raven's Advanced Progressive Matrices (RAPM) was administered at pretest and posttest in counterbalanced order. Subjects were given 10 minutes to solve the 18 items in each set. RAPM has excellent construct validity with respect to fluid intelligence as determined by latent variable analysis (Engle, Tuholski, Laughlin, & Conway, 1999).
- *Working Memory:* the WAIS-IV Working Memory Index (WMI), calculated from forwards/backwards/ordered digit span, mental arithmetic and letter-number sequencing. The latter is optional and was skipped for some fatigued patients.
- *Working Memory:* A computerized Operation Span (Unsworth, Heitz, Schrock, & Engle, 2005) with 3 x span 2-4. Each subject was scored using the partial credit unit scoring method (Conway et al., 2005) which is the average proportion of items recalled in the correct location in each trial.
- *Processing Speed:* the WAIS-IV Processing Speed Index (PSI), calculated from symbol search and digit-symbol coding. PSI and WMI have high internal consistencies and retest reliabilities (Iverson, 2001).

- *Processing speed with inhibition:* 180 trials on a computerized Stroop task of which 20 % were incongruent. Subjects responded verbally to maximize interference (Liotti, Woldorff, Perez, & Mayberg, 2000) while pressing a key to register reaction time. Psychometric properties?

The operation span test and the stroop test were computerized using PsychoPy (Peirce, 2007) version 1.79 and administered in a separate test session.

Statistical models and inferences: Outcome data was modeled in R 3.1.2 as mixed models with main effects of time, treatment and group and their interactions using the lmer 4.1.1 (Bates, Maechler, Bolker and Walker, 2014) and BayesFactor 0.9.10 (Morey & Rouder, 2015) packages. There was a random intercept per subject to account for correlations between repeated measures. Inference was based on model selection between a full model and a null model. The null model was the full model less the fixed effect in question, e.g. the three-way interaction.

Training data was modeled using a power function ($N=k+ax^b$) for task level (N) as a function of time (block number) with random intercepts (k) per subject and random a and b parameters for intervention and group respectively to reflect differences in gain.

p-values from the chi-square statistic of a likelihood ratio tests (henceforth LRT; see Barr, Levy, Scheepers, & Tily, 2013) were reported to comply with current publishing practices. It has long been known that p does not quantify evidence for/against the null and therefore has poor inferential value (Berger & Sellke, 1987; Sellke, Bayarri, Berger, 2001; Wetzels et al., 2011). Bayes factors (BF) with a relatively uninformative Cauchy(0, $\sqrt{(2)}/2$) prior on each covariate were used to quantify the relative evidence for each model (Rouder & Morey, 2012) with the exception of the power function where a more uninformative unit information prior (Wagenmakers, 2007) was used for computation convenience. A Bayes Factor is the odds ratio between two models. For example, BF=5 means that this data shifts the odds 5:1 in favor of the full model and simultaneously shifts the odds 5^{-1}=0.2 for the null.

2.2.2 Results

Training tasks: The progression on the training tasks can be seen in Figure 5 and effect sizes for the training task in Table 3. Block number was used as time unit with a total of between 200 and 240 blocks for completers (20 days x 12 blocks per day). The average of the power fits for individual patients are superposed on the data in Figure 5. The power model was by far preferred to an intercept-only model for all four group x treatment cells ($p_{LRT} << 0.001$, $BF_{BIC} >> 1000$) providing evidence that there was improvement on the training task in all conditions. I.e. the power model is much more than 1000 times more

likely than the intercept model given the data. This is of crucial theoretical importance since it establishes the training as a potential source of transfer to the outcome measures. It was also preferred to a linear model with the same random-effects structure ($p_{LRT} << 0.001$, $BF_{BIC} >> 1000$), thus providing evidence that the data is more likely to have been generated by a power law than a linear law. It is apparent from Figure 5 that the NI group improved much more than the ABI group numerically. This was confirmed by a substantial group-specific testing the random effects of group on a and b ($p_{LRT} << 0.001$, $BF_{BIC} >> 1000$).

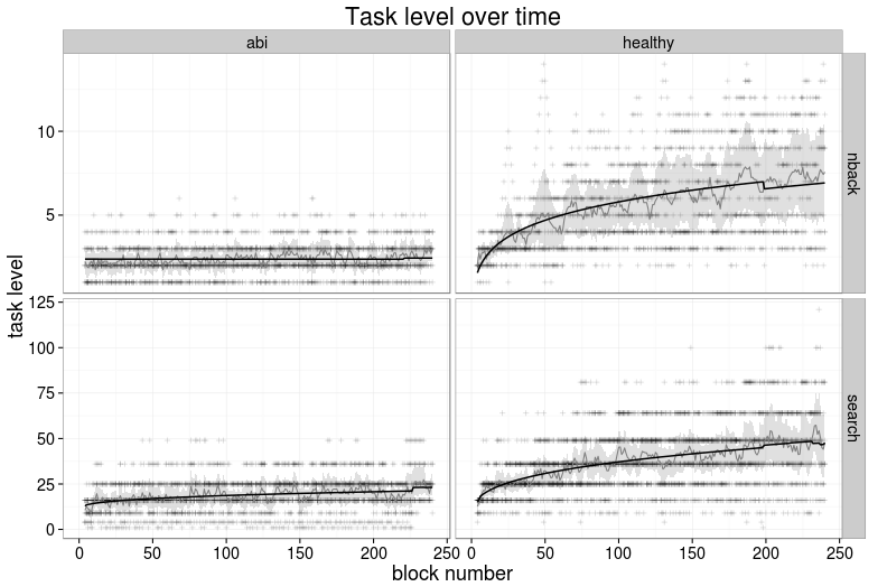

Figure 5: *Level of training task as a function of block number (out of 240 in total) for completers in each group by task. The ABI group was consistently at a lower level and improved little on the training tasks whereas the non-injured group improved on both tasks. The thick black line is the average of the predictions from a fitted power function. The thin gray line and the gray area are the means and 95% bootstrapped confidence intervals for the mean level at each block number. The transparent '+' symbols are the data from individuals which the above represents. Two NI completers improved to N=27 and N=45 on the N-back task and were not plotted. The breaks in the power function after block 200 is caused by subjects who stopped training there. Plots for all individuals are available in supplementary Figure 6.*

	N-back			Visual Search		
group	**Pre**	**Post**	**SMD**	**Pre**	**Post**	**SMD**
ABI	2.1 (1)	2.6 (0.8)	0.45	13.9 (8.1)	22.5 (11.9)	1.06
NI	2.7 (0.8)	7.5 (3.5)	6.11	20.2 (7.7)	48.1 (23.6)	3.58

Table 3: *Pre- and post-test means, standard deviations and standardized differences for task level N. "pre" is computed from the the first 12 blocks starting from the first block where the subject did not increase a level. "post" was is computed from the last 12 blocks. Given the large sample of blocks, even very small effects would trivially would fall under the full model (p_{LRT} << 0.001 and BF_{BIC} >> 1000) so we consider effect sizes is the most informative statistic here.*

Outcome measures. See Table 4 for descriptives and inferences on outcome measures. A difference in effect between the Visual Search training and the N-back training was not supported by the data on any outcome measure, neither in the patient group nor the non-injured group. All Bayes Factors favored the model without the treatment x time interaction with a BF of about 2:1 against the full model. This BF is suggestive but too weak to say anything definitive about the direction of the effect as it still puts around 1/3 posterior probability on the full interaction model. Furthermore, the controlled standardized mean difference ($SMD_c = SMD_{nback} - SMD_{search}$) were in the zero-to-small range (around 0.0 to 0.3) with no signs of a differential effect between single tests or cognitive domains.

Interestingly, the data suggests that the ABI and NI group did not differ in their gains on N-back nor Visual Search on any outcome as revealed by inferences on the group x time x test interaction term (see Table 4). Bayes Factors again favored the null more than the alternative and no interactions were statistically noticeable (p > 5%).

		N-back				Visual search				time x treat			ti x tr x grp	
		PRE		POST		PRE		POST						
	group	mean	sd	mean	sd	mean	sd	mean	sd	d_c		BF	p	BF
WM index	abi	88.00	22.06	92.75	23.29	91.56	16.19	89.67	14.75	0.36	0.14	1.19*		
	healthy	96.78	15.75	101.44	15.39	99.22	18.34	101.22	14.77	0.16	0.43	2.03*	0.47	1.94*
PS index	abi	84.12	13.86	92.00	10.10	83.00	14.00	87.00	14.90	0.29	0.23	1.66*		
	healthy	109.56	15.47	121.00	15.27	107.00	13.87	119.00	18.73	-0.04	0.88	2.35*	0.38	2*
RAPM	abi	3.75	2.66	3.25	2.71	4.11	1.90	4.11	1.36	-0.23	0.59	2.18*		
	healthy	8.89	2.80	10.67	2.50	8.78	4.76	9.89	3.79	0.18	0.66	2.36*	0.51	2.22*
OSPAN	abi	74.69	21.94	74.07	15.16	71.30	9.15	70.52	19.80	0.01	0.99	2.00*		
	healthy	59.26	16.85	64.40	20.30	65.02	10.35	69.75	12.42	0.03	0.94	2.52*	0.98	1.53*
logstroop	abi	0.42	0.13	0.35	0.11	0.42	0.03	0.39	0.05	-0.52	0.44	1.81*		
	healthy	0.31	0.12	0.28	0.12	0.26	0.07	0.25	0.12	-0.17	0.70	2.12*	0.84	1.95*

Table 4. Means, standard deviations and univariate inferential results for all outcome measures. WM=Working Memory, PS=Processing Speed. Stroop times are given in the logarithm of milliseconds since they were approximately log-normal. SMD_c is the controlled standardized effect size, computed using pretest standard deviation. p-values are from the likelihood-ratio test of the critical interaction and BF_g=Bayes Factor with g-priors on regression coefficients (Rouder & Morey, 2012). * asterix indicates that the BF_g is in favor if the null ($1/BF_g$). *Briefly, "time x treat" answers the question: "is the tasks associated with different gains for this group?" and "time x treat x group" answers the question "is the controlled N-back gain different between groups?" Results for individual subtests and for stroop reaction time is available in the supplementary materials*

2.2.3 Discussion

We observed an improvement on the training task but no transfer to untrained tasks. Therefore the Cicerone et al. (2002) findings were not replicated. This should be regarded as a conceptual replication of the N-back part of the Cicerone intervention since there are a number of design differences, including that Cicerone et al. provided the intervention in person, used different materials and different outcome measures.

With respect to the N-back literature on non-injured adults, these results fail to replicate findings of positive transfer (Jaeggi, Buschkueh, Jonides & Perrig, 2008; Jaeggi et al., 2010) but are in line with several null findings (Jaeggi, Buschkuehl, Jonides & Shah, 2011; Redick, Shipstead & Harrison, 2012; Chooi & Thompson, 2012). The former studies used passive control groups while the latter used active control groups. It is possible that the initial positive results were caused by confounding nonspecific factors, such as expectation and motivation, which cannot be attributed to the task itself. If that is the case, these studies do not constitute evidence that the training task per se caused the observed effects.

Convergence: computerized training yield specific effects. N-back and Visual Search training did not differentially improve performance on neuropsychological tests which are thought to reflect working memory, processing speed or fluid intelligence. Since previous research has shown that N-back performance reflects working memory and that VS does not ([18] and we observe no selective effect of N-back improvement on working memory measures, we conclude that subjects developed specific strategies to solve the N-back task during the course of training.

How specific? Answer: So specific that N-back training on digits did not transfer to digit span (part of WAIS Working Memory Index), N-back training on letters did not transfer to Operation Span letter recall, Visual Search training did not transfer to Symbol Search (part of WAIS, see supplementary) and N-back training on locations in a 3 x 3 grid did not transfer to the RAPM 3 x 3 grid. It is clear that these data pose problems for the naïve view that abstract cognitive abilities were acquired during the training and even the view that something intermediate was trained, such as recalling sequential items or scanning visual grids. Instead, we believe that computerized training follows the well-known learning principle that the efficiency of the decoding/use of a skill is proportional to the similarity of the decoding context to the encoding/learning context of this skill (Perkins & Salomon, 1989; Tulving & Thomson, 1973). The repetitive nature of computerized training make up a highly stable context and thus the N-back and VS skills become “locked” to this context in a way that does not generalize to a neuropsychological test setting or even the computerized tests.

This is evidence that the development of specific strategies that do not transfer to untrained tasks might be a possible point of convergence between a non-injured and a brain injured

population. Although specific improvement may seem unflattering compared to generalized improvement, it could actually be thought of as a very efficient information processing strategy where the cheap local strategies are preferred to the slow and costly high-level cognition (Clark, 2013). As such, the tendency and ability to develop specific strategies could be regarded as a property of a healthy cognitive system.

Divergence: the formation of specific skills. The non-injured group improved 2.5-5.5 SMD more on the training tasks than the ABI group. We interpret the training data to reflect an impaired ability to form specific strategies in the ABI group. One explanation for this observation is that well-functioning domain-general cognition is necessary for the effective formation of specific strategies. However, 5 out of 17 ABI patients had a baseline Working Memory Index score over 100 and two patients had a Processing Speed Index score over 100. But neither of these improved nearly as much on the training task as the average person in the NI group, thus discrediting this hypothesis.

The group differences in etiologies could be confounded by the age difference. However, Dahlin et al. (2008) did computerized training on non-injured young and elderly adults and found no difference in gain on the training task. A meta-analysis on 25 computer-based parallel-groups RCT on brain injured patients similarly found no effect of age on improvement (Lindeløv, in preparation). Thus both within-study and between-study evidence discredits age as the sole explanation for the observed discrepancy between the ABI and NI group.

This leaves us in a limbo with no single candidate explanation for the observed difference between groups. There is a vast literature on the topic of specific learning impairments following acquired brain injury (Vanderploeg, Crowell, Curtiss, 2001; Schmitter-Edgecombe, 2006). However, it has almost exclusively investigated verbal and motor learning within single sessions. Four weeks of training on the N-back task and the Visual Search tasks do not readily subsume under these categories so the present study may contribute new evidence. It is up to future studies to narrow in on the mechanisms driving this effect. To support this effort, we encourage authors to report and interpret training data explicitly when doing this kind of study.

Limitations: The final sample size per condition is small even though a total of 78 participants initiated the training. The present study should not be considered as basis for clinical guidelines but rather as preliminary evidence which puts a few ideas about the mechanisms underlying computer based training on the table. Even then, is not smaller than the two most cited studies on the topic (Westerberg et al., 2007; Sturm, Willmes, Orgass & Hartje, 1997). The sample size was influenced by a large dropout rate of 50% which was not biased with respect to age, gender, baseline scores or (for patients) FIM score. The dropout rate may be informative about the expected adherence to fully self-initiated training without a monetary reward for NI and ABI subjects alike. The optimal evidence support for

the conclusions above would have been obtained if the ABI and NI group had been matched on all nuisance parameters such as age, education and socioeconomic status.

Future directions: We suggest that future research on computerized cognitive rehabilitation may progress along two different routes: (1) Prevent specific learning. This is not easy. Simply training on a large array of different tasks may not be sufficient as demonstrated by several null findings who did this (Middleton, Lambert & Seggar, 1991; Chen et al., 1997). A true context-breaking intervention would constantly present novel problems, shifts between devices, change colors, be trained at different locations etc. We expect that this approach is too chaotic to be feasible with a patient population. Alternatively, (2) exploit the context-specific effects and make the training task as similar to the transfer target as possible, i.e. practice reading television subtitles, doing mental arithmetic on shopping costs etc. For example, Yip and Man (2013) successfully improved real-life shopping performance after training in a matching virtual reality environment. This is a much less ambitious target than high-level cognition but may also be more realistic.

2.2.4 Acknowledgements

The authors would like to thank Hammel Neurorehabilitation Center for cooperation and Lars Evald for helpful comments on an early draft of this paper. This work was supported by a grant from Hospitalenhed MIDT.

2.2.5 Supplementary training data

Figure 6 shows training data for all 78 individual subjects together with predictions from the linear mixed model which as an R formula can be represented as *log(N) ~ block_number * group * treatment + (1|id)*. This model is quite simple: six single-valued parameters common to all participants and then a subject-specific random effect. That is a reduction from 8834 data points to 43 parameters (37 subjects + 6 fixed effects). We find that the model represents the individuals well. Another option would be to fit a quadratic time-trend instead of log-transforming N.

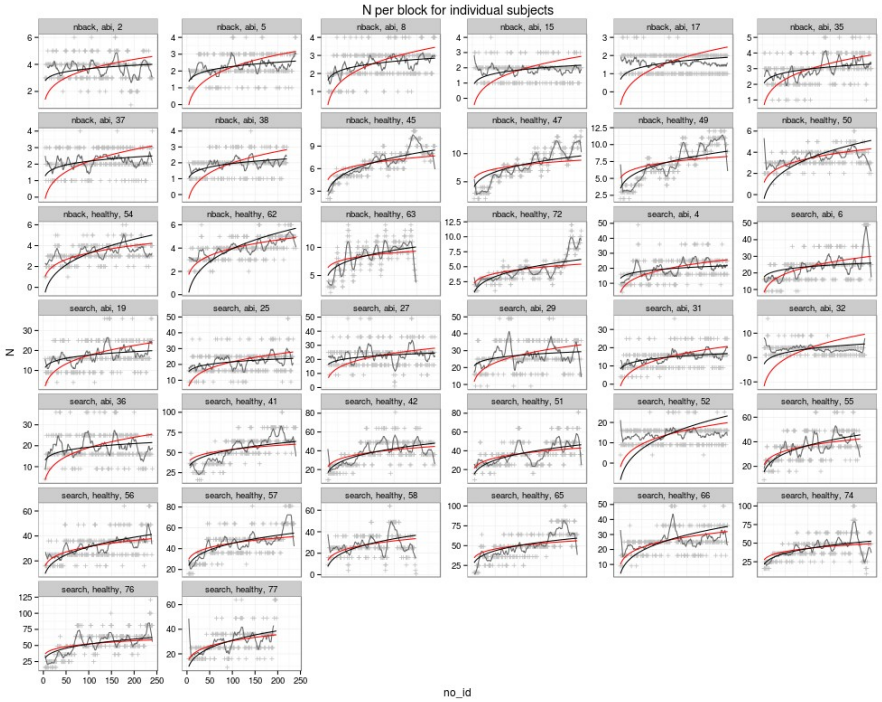

Figure 6: *N as a function of block number during training for all 38 completers. Plusses are individual data points, the irregular line is a 12-block running mean. The black line is the fit by a full power function (with a random a, b, and k for each subject) and the red line is the fit by the null model of no difference between group gains (only random k). The full model has a better fit.*

Although the fit of a model with a fixed training effect yield good fits, an inspection of the distribution of the first and last blocks reveals a nuance to the story. Figure 7 shows the distribution of the first and last 12 blocks for each group x treatment cell. For the ABI group where there is little improvement, the data are well represented by a shift in mean and equal distributions. However, for the NI group, the distribution widens considerably, reflecting a variability in the individual's improvement. Here a fixed-effect-and-variance view does not do justice to the data.

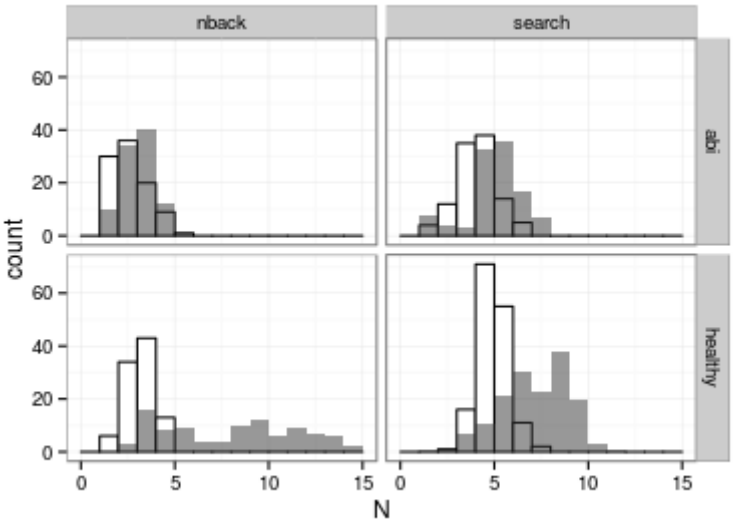

Figure 7: *distribution of task levels for the first (unfilled histogram) and last (filled histogram) 12 blocks by completers.*

2.2.6 Supplementary outcome data

Table 5 includes the subtests that make up the WAIS working memory index and processing speed index. It also includes the absolute mean reaction time on the stroop task.

		N-back				Visual search				time x treat			ti x tr x grp	
		PRE		POST		PRE		POST						
	group	mean	sd	mean	sd	mean	sd	mean	sd	d_c	p_LRT	BF	p	BF
digit span	abi	7.25	4.53	8.38	4.17	7.78	3.19	8.22	2.99	0.18	0.41	1.96*		
	healthy	10.44	3.43	10.67	2.83	10.50	2.39	9.88	2.30	0.29	0.32	1.8*	0.89	2.19*
arithmetic	abi	8.75	3.69	9.88	4.02	9.88	3.27	9.75	2.76	0.37	0.14	1.16*		
	healthy	9.56	2.70	10.00	2.50	9.75	2.87	10.25	2.87	-0.02	0.92	2.46*	0.02	1.35*
let-num-seq	abi	10.67	5.09	12.00	4.10	9.38	2.33	9.88	2.10	0.23	0.58	2.12*		
	healthy	8.56	2.46	10.00	3.16	9.50	3.02	9.62	3.11	0.49	0.09	1.07	0.76	2.14*
WM index	abi	88.00	22.06	92.75	23.29	91.56	16.19	89.67	14.75	0.36	0.14	1.19*		
	healthy	96.78	15.75	101.44	15.39	99.22	18.34	101.22	14.77	0.16	0.43	2.03*	0.47	1.94*
symb-search	abi	7.12	2.03	8.25	1.83	7.00	2.40	7.56	2.60	0.26	0.36	1.87*		
	healthy	11.56	2.55	13.00	4.30	11.88	3.14	13.62	3.70	-0.11	0.81	2.37*	0.53	2.31*
symb-code	abi	6.88	3.23	8.12	2.53	7.00	2.62	8.25	2.43	0.00	1.00	2.36*		
	healthy	11.89	3.06	13.89	2.98	10.62	3.46	12.75	3.85	-0.04	0.88	2.28*	0.91	2.44*
PS index	abi	84.12	13.86	92.00	10.10	83.00	14.00	87.00	14.90	0.29	0.23	1.66*		
	healthy	109.56	15.47	121.00	15.27	107.00	13.87	119.00	18.73	-0.04	0.88	2.35*	0.38	2*
rapm	abi	3.75	2.66	3.25	2.71	4.11	1.90	4.11	1.36	-0.23	0.59	2.18*		
	healthy	8.89	2.80	10.67	2.50	8.78	4.76	9.89	3.79	0.18	0.66	2.36*	0.51	2.22*
ospan	abi	74.69	21.94	74.07	15.16	71.30	9.15	70.52	19.80	0.01	0.99	2*		
	healthy	59.26	16.85	64.40	20.30	65.02	10.35	69.75	12.42	0.03	0.94	2.52*	0.98	1.53*
logstroop	abi	0.42	0.13	0.35	0.11	0.42	0.03	0.39	0.05	-0.52	0.44	1.81*		
	healthy	0.31	0.12	0.28	0.12	0.26	0.07	0.25	0.12	-0.17	0.7	2.12*	0.84	1.95*
logstroopRT	abi	6.89	0.04	6.94	0.09	6.97	0.11	6.95	0.04	0.7	0.33	1.37*		
	healthy	6.52	0.27	6.41	0.29	6.66	0.19	6.54	0.11	0.03	0.93	2.23*	0.65	2.02*

Table 5: *pre- and post-test means and standard deviations, controlled effect size and inferences for all outcome measures. See table text for Table 3 for further details.*

2.3 Epilogue: Is the N-back task a working memory task throughout the experiment?

The claim that brain injured patients have an impaired ability to learn specific strategies deserve more discussion and scrutiny than space allowed for in the paper. However, that discussion will be postponed to after the next paper, since they inform each other. Here, I want to expand on what this finding implies about the underlying cognitive change during training. This was the line of thought that lead me to do a review and meta-analysis and therefore also serves as an introduction to the review paper.

2.3.1 Construct validity of the N-back training task

The N-back task is classically thought of as a working memory task but it shares little variance with complex span tasks, such as Operation Span, even though both are considered to be working memory tests. Kane et al. (2007) found that OSPAN explained around 11% of the variance in N-back while Redick et al. (2013) found it to be 4%. This surprisingly weak relationship is also observed in intervention studies with non-injured participants where N-back performance increases with training while complex span performance remained stable in the same subjects (Jaeggi et al., 2010; Redick, 2012). This is a reminder to be cautious when assigning abstract "cognitive labels" to particular tasks. As I have argued, it is more accurate to say that N-back performance pertains to working memory functioning than to equate the two, as is all too common (e.g. in the titles of Jaeggi et al., 2008; 2010; Redick et al., 2012; Chooi et al., 2012).

In our study, we also find an increase in N-back performance and a relatively unchanged performance on the operation span task and Raven's matrices from pretest to posttest. Interestingly, Kane et al. (2007) found that the N-back task covaried with fluid intelligence as assessed by Raven's matrices (though with a medium sized r=0.33), but we did not observe an improvement in RAPM even though N-back performance improved. This dissociation is most striking in healthy subjects since they improve the most on the N-back task. If performance on N-back and RAPM reflected a bottleneck in the same underlying construct, improvement in N-back would not be possible without a widening of this bottleneck which in turn should cause better performance on RAPM. This is at odds with our findings. This conflict between findings of cross-sectional studies and longitudinal studies shows itself in many experiments. To mention a few, Chooi et al. (2012) and Redick et al (2012) observed improvement on the N-back training, but not on RAPM outcome.

Harrison et al. (2013) observed improvements on complex span training but not on RAPM outcome.

I can think of two possible theoretical explanations of this.

2.3.2 Theoretical accounts of near-transfer

#1: Low-level strategy offloads high-level cognition. It may be that subjects develop specific strategies to solve the task and use these to leverage the load on higher cognitive processes. Such an explanation fits well with many learning theories, e.g. the ACT-R framework (Singley & Anderson, 1989) and the free energy account of information processing in the brain (Friston, 2010; Clark, 2013). High-level processing is expensive and low-level processing is cheap. So subjects may draw on high-level cognitive functions initially but eventually develop specific strategies at which point the lower layers can efficiently handle the task without loading the higher layers. If this shift happens at an early stage, say during the first two or three training sessions, the high-level function is never really "trained" extensively[22] - or at least it has laid relatively dormant for a long time at post-test. In other words, the training task changes construct validity during the training. An N-back task may initially share a great deal of variance with other "working memory tasks" but almost exclusively share variance with itself in the end of the training. If this is true, this effective strategy of using as few mental resources as possible is the very principle that underlies the dominance of near-transfer over far-transfer, as suggested by these theories.

#2: High-level cognition is illusory. An alternative explanation is that there is no unitary or causally intertwined high-level cognitive processes in the first place. According to this view, it *appears* that different behaviors expresses the same underlying capacity limits (a view most forcefully promoted by Miller, 1956 and Cowan, 2000, 2005) while the underlying cognitive architecture is not causally related at any given time. Andy Clark formalized this line of thought under the heading "Microcognition" or "microfunctionalism" (Clark, 1991), according to which cognition is atomic at its core, composed of very simple independent functional units. These units can, however, interact in ways that establishes higher-order (or "emergent") functionality.

So the correlations found in cross-sectional studies does not necessarily imply that the underlying cognition is causally related. According to this view, the reason that N-back performance can increase while RAPM does not, is that they were never causally related. The observed correlations between e.g. working memory behaviors may be caused by a third factor. One prime candidate as a third factor is experience. All mental faculties are formed by the same evidence base - the lifetime experiences of the subject - and this causes some convergence in aptitudes. When subjected to computer-based training, a small set of

22 This is the muscle metaphor of cognitive improvement, which I will return to in the discussion

these “microfunctions” are selectively exposed to experiences and thereby they are being de-correlated from the other “microfunctions” which are not brought to bear on the training task. As a result we observe a near-transfer effect.

Evaluation of the explanations: I see no easy way that these two explanations could be distinguished empirically from behavioral observations since they make the same predictions[23]. However, the prospect of *any* of them being a more accurate account of learning and transfer in training experiments is interesting in and of itself. A first step towards informing *that* distinction would be for authors to report the actual training data.

2.3.3 Examples and implications

An important implication, if these speculations hold, is that the psychometric properties of the training task changes during training. Specifically, trained behaviors become optimized for the training task. If one extrapolated from this narrow set of abilities to the whole domain, one would overestimate the effect. As an extreme case, Maguire et al. (2002) investigated ten world memory experts compared to matched controls and found no important differences in performance on classical intelligence and memory tests. Maguire et al. attributed differences in memory performance to the use of specific strategies that can be applied in a very narrow set of situations (see also Ericcson, 2003). It would be a gross error to extrapolate from the memory experts’ specific memory skills to their general memory abilities. Anecdotally, I have experienced this myself. I reach a 15-back level on the N-back task, a 7-back level on a dual N-back task, and a 3-back level on a quadro N-back task. To my great regret, I did not notice any improvements in my cognitive abilities on other materials, e.g. the ability to concentrate on academic papers or to comprehend complex arguments. I later learned to recite pi with 130 digits and the natural number with around 60 digits. This did not enhance my ability to learn phone numbers and I did not spontaneously see fragments of pi decimals in the world around me. For both N-back and digits, the ability to do well was tightly locked to the exact material I had been training on. The same may be true for the participants in our N-back study and in the many other studies cited in this section.

The above considerations about the potentially changing role of the training task and its implications for transfer became apparent to me only after having conducted the N-back studies and seeing the results. If the above is true, many studies (the N-back study included) have confounded stimulus-specific effects with proper transfer effects. That was the point of departure for the review which takes methodological confounds into account.

23 I have thought about this for three years now to no avail.

2.4 Computer-based cognitive rehabilitation following acquired brain injury: does the evidence support the claims?

Four necessary criteria for the claim that a behavioral interventions improves a cognitive function are presented. A claim is supported if (1) the outcome measures represents the targeted cognitive domain. This entails that (2) outcome measures are dissimilar to the training task. (3) Changes in outcome measures must be relatively consistent. These changes can only be attributed to the intervention if (4) the intervention is isolated using an active control group.

A review on computer-based cognitive rehabilitation included 30 studies with a total of 42 parallel-groups treatment-control contrasts. At least 27 contrasts studies were categorized as false positives according to these criteria. In other words, the publications claimed to find cognitive transfer effects but did not provide the necessary evidence. At most three contrasts were categorized as a true positives. There were 12 true negatives (null findings).

A meta-analysis found no effect on non-trained outcome measures (SMD=0.05, 95% CI=-0.10 to 0.20), i.e. no transfer. The observed positive effects could be attributed to low-level similarities between training and outcome tasks (add 0.24 SMD) and nonspecific effects which go untested in passive control groups (add 0.23 SMD).

Thus the very literature which is cited in support of cognitive transfer turns out to provide strong evidence for the opposite view: computer-based cognitive rehabilitation does not improve high-level cognitive functioning.

Systematic reviews and meta-analyses frequently classify the quality of Randomized Controlled Trials using the Jadad scale (Jadad, Moore, Carroll, et al., 1996). This scale assigns a 0-5 score to individual studies based on their procedures for randomization, blinding and handling of dropouts. Outcome measures for "high-quality" studies are classified into mental categories (e.g. "attention", "executive functions", "memory" etc.) and compared across studies. By doing so, however, a number of implicit theoretical assumptions are silently accepted and issues neglected. For example, several studies have concluded that attention improved following computer-based cognitive training although

improvement was only observed on half or fewer measures of attention relative to control groups (e.g. Sturm, Willmes, Bernt & Wolfgang, 1997; Sturm, Fimm, Cantagallo, et al. 2003; Gray, Robertson, Pentland & Anderson, 1992). This is hardly compatible with the view that attention improved overall. Other studies have used the training task itself as an outcome measure (e.g. Prokopenko, Mozheyko, Petrova, et al., 2013; Man, Soong, Tam, Hui-Chan, 2006; Stablum, Umiltà, Mogentale, Carlan, & Guerrini, 2000) and interpreted improvement as if it reflected a general cognitive capacity, although these results are expected under the hypothesis of near transfer and even zero transfer.

Still, these studies are cited as were they evidence in support of a generalized cognitive improvement (e.g. in Cha & Kim, 2013) even though the results are supportive of the opposite hypothesis: that any cognitive improvement is fairly specific to the trained task. The interpretations laid out in the papers are silently accepted even when they are at odds with the very method and data they are based on.

While the Jadad scale may be a good general-purpose tool to evaluate bias in empirical studies, I will argue that additional criteria are necessary for behavioral interventions targeted at cognitive improvement. The next section motivates this necessity and the following section presents four such additional criteria.

2.4.1 Training effects and transfer effects

Under what circumstances may claims like the following be supported?

> *Training on [task] improved [cognitive function]*

… where the cognitive function could be a high-level construct such as attention, working memory, prospective memory, executive functions, selective attention and the like.

To answer this question, it is important to distinguish between training effects and transfer effects (Barnett & Ceci, 2002). Loosely, a training effect is a change in aptitude which is closely bound to the material and context in which it was acquired. It is sometimes called narrow learning, near-transfer, skill learning or low-level strategies. A transfer effect is any change in aptitude which extends beyond the training effect, i.e. generalizes to new material and new situations. It is sometimes called wide learning, far-transfer, remediation or generalized learning. Since cognitive functions usually take a wide variety of stimuli and contexts as input, an improvement of a cognitive function by means of a limited set of stimuli and context(s) would require a transfer effect rather than a training effect. Therefore, one has to present convincing evidence for a transfer effect to make the above claim.

Naturally, transfer effects are desirable in a rehabilitation setting where the goal is to improve patients' functioning in the community rather than their performance on neuropsychological tests in a rehabilitation setting. Unfortunately, numerous studies on learning have supported the view that learning and recall tends to be highly specific to the conjunction of stimuli in the learning situation, starting with Edwin Thorndike and colleagues more than 100 years ago (Thorndike & Woodworth, 1901 later extended by Singley & Anderson, 1989). It has later been described in cognitive psychology as the principle of encoding-decoding specificity (Tulving & Thomson, 1973) and in educational psychology as the context-specificity of learning (Perkins & Salomon, 1989). Within brain injury rehabilitation it is closely related to the distinction between remediation and compensation (Dirette et al., 1999). Here too, many have advocated the view that interventions should be very specific to the behavior one wish to improve (Wilson, Baddeley, Evans & Shiel, 1994; Park & Ingles, 2001)

In addition to this, the method of computer-based training is the "drill and practice" strategy which is particularly known to form specific associations that are effective for rote learning of explicit facts but not for thinking skills (Perkins & Salomon, 1989; Tournaki, 2003; McGurk, Twamley, Sitzer, McHugo, & Mueser, 2007). It may even hinder the development of such abilities by over-training more specific solutions, leading to negative transfer when subjects apply overlearned behaviors in situations where they are sub-optimal (Singley & Anderson, 1989).

Thus I argue that the broader theoretical and empirical literature on human learning and rehabilitation leaves the burden of proof on demonstrating transfer effects. But what would it take to establish such a support for transfer effects in a study on computer-based cognitive training? I propose four necessary criteria to this end in the next section. In the following sections, a review applies these criteria to single studies and a meta-analysis synthesize evidence across studies.

2.4.2 Necessary methodological criteria

In this section I will argue that it is necessary to (1) have a test battery that represents the full transfer domain, which entails (2) having outcome measures that are dissimilar to the training task so that potential low-level improvements are not interpreted as proper cognitive transfer. (3) Outcomes satisfying the first two criteria should show a consistent effect, which (4) can only be attributed to the computer-based aspect of the intervention if contrasted with a parallel active control group. One may regard this section as an explicitation of how the experimental method would flesh out when establishing the effect of training (independent variable) on a cognitive function (dependent variable). Satisfying Representativeness, Dissimilarity, and Consistency increases the experimental validity of the dependent variable(s) and satisfying the active control criterion increases the

unambiguity of the independent variable.

The criteria will be presented in a fairly abstract way here but they are applied to 30 publications in the two following sections where they can be seen in action.

Representativeness: outcome measures should exhaust the transfer domain

The aptitude of a cognitive function does per definition apply to all stimuli and processes that depend on it. For example, if attention is improved, you would predict that all behavior which is mediated or moderated by attention should be affected to the degree that attention is involved. Therefore, a perfect support for the claim that attention improved would be to address all of these behaviors and all should be affected in a way that is consistent with an underlying improvement of the attention-specific involvement in each measure. This includes a wide array of stimuli for each sensory modality (visual, auditory, tactile, olfactory etc) as well as interoceptive signals.

Such an exhaustive mapping of behavior is of course too comprehensive or even impossible to do in practice. However, we can turn a logical argument into a probabilistic argument by doing inference on representative samples. Just like we sample individuals because it is unfeasible to test a full population, we can sample behaviors from the transfer domain. But just like a single subject is insufficient to study the nature of human behavior in general, a single outcome measure is insufficient to make claims about changes in domain-general cognitive functions. Even two or three different subjects/tests seem like a very inadequate sample although it is certainly much better than just one. Staying with the analogy, you should be careful about generalizing to all of humanity if you only study subjects from one culture, e.g. a western university (Henrich, Heine, & Norenzayan, 2010). Similarly, if the aptitude of attention is measured using only visual tasks, you may make valid claims about visual attention but should be careful about making inferences on the aptitude of attention in untested modalities.

However, we can administer fewer tests than the analogy to sampling of subjects might imply, because we usually do not know the properties of recruited individuals in advance but we do have knowledge about many tests. Using varieties of structural equation modeling (SEM, e.g. factor analysis, latent variable analysis and simple correlation) we have approximated to what degree performance on particular tasks are representative of each other. For example, Ravens Advanced Progressive Matrices explains around 50 - 80 % of the variance shared by measures of fluid intelligence and Operation Span explains around 60 % of the variance shared by complex span measures of working memory (Engle, Tuholski, Laughlin, & Conway, 1999; Kane, Hambrick, Tuholski, et al., 2004; Kane, Hambrick, & Conway, 2005). SEM could be an important tool to select an experimentally feasible number of outcomes that best represents the full transfer domain.

In summary:

> *Representativeness recommendation: Administer multiple outcome measures with high construct validity for the cognitive domain of interest.*

Dissimilarity: outcome measures should be dissimilar to the training task(s)

The training task is often a part of the transfer domain. I.e. to improve attention, you train a task which strains attention with the aim of improving it like a muscle. To improve working memory you train a task that strains working memory. The same is true for problem solving, reasoning, prospective memory etc.

I have argued that learning tends to be stimulus-specific. As a consequence, as subjects improve on the training task, it is very likely to become a local aptitude maximum in the transfer domain. Consequently, the performance on the training task becomes non-representative of the transfer-domain and cannot be used as evidence on equal footing with non-trained tests since it is a priori very likely to overestimate the underlying effect.

This is also true of outcome measures which are similar but not identical to the training task. For example, if digit span is trained to improve memory and letter span is used as outcome measure, and both are delivered in the same modality, at the same pace, and at similar span lengths, there could be some carry-over of specific optimizations for these similarities which are peripheral to a general memory capacity. Again, the consequence is that the letter span becomes unrepresentative of the memory transfer domain.

More generally, a measure within the transfer domain becomes positively biased to the extent that it overlaps with the training domain. Its representativeness is "contaminated" by the specific strategies developed during training. Therefore, the training domain becomes a "blind spot" in the transfer domain, since stimulus-specific effects cannot be disentangled from effects of the targeted cognitive function. So:

> *Dissimilarity recommendation: minimize stimulus-specific overlap between the training task(s) and the transfer task(s) of interest. Evaluate the extent to which any remaining overlap accounts for observed transfer.*

Consistency: changes in outcome measures must be (largely) consistent

Obviously, merely designing an appropriate test battery (satisfying Representativeness and Dissimilarity) is insufficient to establish that a high-level cognitive change occurred. You need data. Ideally outcomes should all improve, all deteriorate or all maintain status quo. In the case that just one outcome measure with 100 % sensitivity to the underlying construct does not change in the same direction as the other outcome measures, the claim that the underlying construct changed as a whole is falsified, no matter how many other tests were administered.

This criterion can be relaxed since there is no cognitive test with 100 % sensitivity or selectivity and therefore any deviation from an overall pattern could be due to influences (noise) from other sources than the targeted cognitive process. However, since the tests ought to have high validity with respect to the construct (cf. the Representativeness and Dissimilarity criteria), any such non-consistent outcomes poses problems for the claim that the construct improved per se and should be interpreted explicitly.

> *Consistency recommendation: Only generalize findings to properties that the improved outcome measures have in common and which does not overlap with the properties of non-improved outcome measures.*

Active control: isolate the cause using a parallel active control group

Satisfying Representativeness, Dissimilarity and consistency is necessary to establish that a high-level cognitive change occurred. However, to establish what caused this cognitive change, various influences must be factored out using a control condition. Such influences usually include retest effects, expectancy effects (placebo) and nonspecific effects of training anything at all (e.g. repeatedly engaging in a cognitively demanding activity). The rationale for using different kinds of control groups is well known but a short explicitation for some common control groups is warranted for the present context.

Given proper stratification and that Representativeness, Dissimilarity and Consistency is satisfied, here is what may be concluded using each control group:

- **Passive**: A mixture of training on [task], expectancy effects and nonspecific effects improved [cognitive function]. The contribution of each is unknown.
- **Treatment as usual**; A mixture of training on [task] and nonspecific effects improved [cognitive function]. The contribution of each was untested but we think that some nonspecific effects are likely controlled for relative to a passive control.
- **Active**: Training on [task] improved [cognitive function].

… where the active control condition would ideally be matched on all nuisance parameters (e.g. motivation, duration and intensity), so that only the hypothesized "active ingredient" differs between groups. That "active ingredient" is usually a particular task or suite of tasks.

Design-wise, I will argue that the only proper controlled design for drill-and-practice rehabilitation is a parallel control group drawn from the same sample as the treatment group. I do this by excluding the two alternatives[24]. First, multiple-baseline designs do not control for retest effects during the treatment-phase and they do not control for expectancy effects and nonspecific effects. The second design is cross-over, which consists of a treatment-control arm and a control-treatment arm. The first phase of a cross-over design is a parallel groups design. But issues arise in the second phase: since the very purpose of rehabilitative interventions is to achieve long-term cognitive changes in the patients, there is no wash-out period so the "control" condition in the treatment-control arm is not comparable to the control condition in the control-treatment arm. Consequently, phase 1, which is a parallel-groups design, is the only properly controlled part of a cross-over design for studies on rehabilitation and training.

To sum up in the last recommendation:

> *Active Control recommendation: use an active control group to establish the training task(s) as the independent variable.*

24 Another type of control is the intermingling of test and control trials on a second-by-second basis, often in basic cognitive science. This kind of control is never used in treatment studies for the same reasons as cross-over studies are inappropriate: it is unsuitable for independent variables which cause a learning effect on the observed variables.

2.4.3 Criteria-based review: does the evidence support the claims?

In this section, the four criteria are applied to studies on computerized cognitive rehabilitation of patients with acquired brain injury. The purpose of this qualitative analysis is to characterize how many of the claims in the literature are substantiated by method and design factors which remain unappraised under the Jadad criteria.

Search strategy

Publications were included in the review if they (1) were predominantly computer-based, (2) targeted higher cognitive functions, which was defined as cross-modal functions. E.g. studies targeting attention are included but those targeting visual attention are not. (3) Used a parallel control group, (4) included patients with acquired brain injury but without selecting for a particular deficit, i.e. neglect, hemianopia, aphasia etc. The underlying purpose of these the inclusion criteria was to homogenize the selection of studies. Details on these criteria as well as search terms are provided in the supplementary A.

PubMed, PsycINFO and Google Scholar was searched for relevant publications. The identification of the included studies are illustrated in Figure 8. References of all papers assessed for eligibility were scanned for further sources as were earlier reviews (Cha & Kim, 2013; Chung, Pollock, Campbell, Durward, & Hagen, 2013; Leclercq & Sturm, 2002; Loetscher & Lincoln, 2013; Poulin, Korner-Bitensky, Dawson, & Bherer, 2012; Rees, Marshall, Hartridge, Mackie, & Weiser, 2007; Rohling, Faust, Beverly, & Demakis, 2009; Xu, Ren, Prakash, Kumar, & Vijayadas, 2013; Xi, 2012). Lastly, publication lists for commercial rehabilitation software was scanned. Two studies had insufficient data for computation of effect sizes (Wood & Fussey, 1987; Sturm et al., 2003) and information about the intervention was too sparse for two studies (De Luca, Calabrò, Gervasi, et al., 2014[25]; Röhring, Kulke, Reulbach, Peetz, & Schupp, 2004, see Table 6). Therefore, the review includes 30 studies and the meta-analysis is based on 26 studies.

Authors were contacted for clarification and nonpublished information/data where needed. Unpublished data and details were obtained for 16 publications, 11 of which were included. Requests to the authors of 15 included publications were not answered. It was not possible to find functional contact information for 8 publications.

25 Per personal communication with authors. Clinical staff did the intervention and the authors do not know what they did.

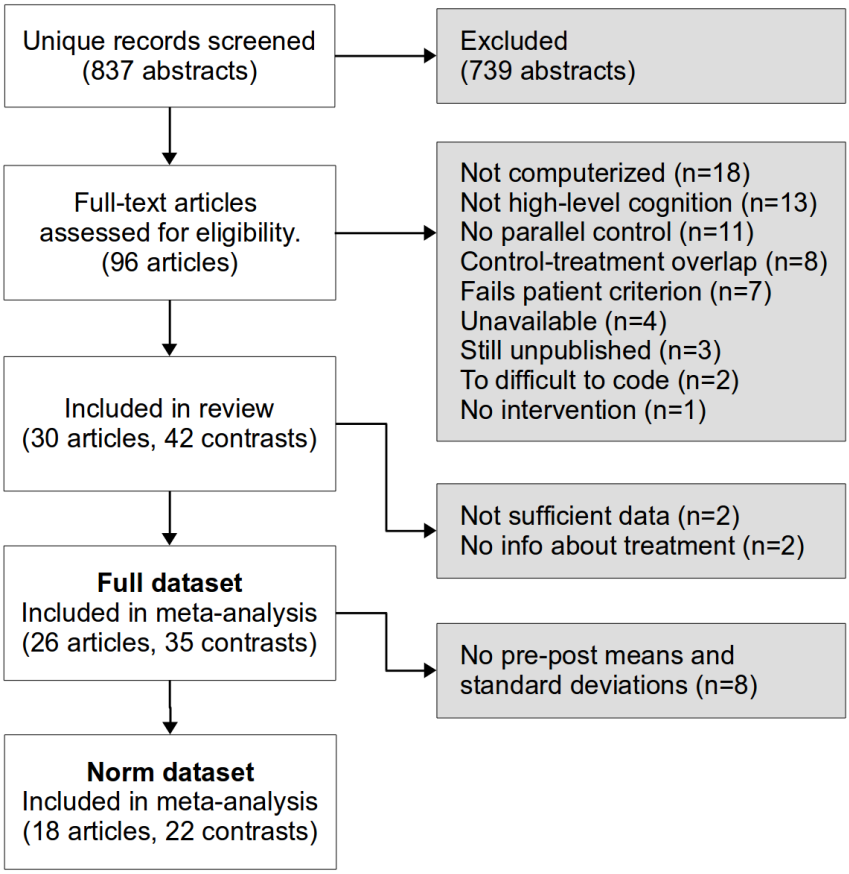

Figure 8. *Flowchart of study inclusion and the studies that went into the review, the full dataset and the norm dataset respectively.*

Outcome extraction and scoring

Only performance-based outcomes were included - typically administered by a neuropsychologist or a computer. Observational scores (e.g. Functional Independence Measure) and self-evaluation questionnaires (e.g. SF-36) were not included in order to homogenize the compared effect sizes and since they are seldom thought to be as direct measures of the aptitude of specific cognitive functions.

Only outcomes which were targeted in the training were included. For example, Yip and Man (2013) trained prospective memory so only the four "prospective memory" outcomes out of 11 outcome measures in total were included. Many training programs simply targeted cognition in general and for these, all cognitive performance-based cognitive outcome measures were scored as targeted.

A total of 298 targeted outcome measures were included and sufficient information about the tests and intervention to judge the training-outcome similarity was obtained for 277 of these (c.f. the Dissimilarity criterion). Each outcome was scored in one of three categories which I hypothesized would show a ranked effect size (large-to-small):

- SAME: the training task and the outcome are essentially identical. Response modalities were allowed to differ. Typically a computer would use keyboard / mouse whereas a testing session uses verbal answers, pointing or manual manipulations of objects.
- SIMILAR: There is an overlap in stimuli and/or structure which could be explained by specific strategies that only applies to the training domain.
- DISSIMILAR: neither SAME or SIMILAR, i.e. there is no obvious low-level explanation for the observed effect on the outcome measure.

When the training task(s) was underspecified further information was sought from authors, from publicly available information material and from other publications that used the same training task(s). Scoring outcomes on these categories is elaborated in the supplementary A and the supplementary dataset provides justifications for every single outcome.

Results

The 30 studies are listed in Table 6 and Table 7. Table 6 contains information about samples and interventions while Table 7 contains information relevant for the four criteria.

The studies cover 1086 patients who were a weighted average of 47.3 years old and a weighted median of 7.5 months post-injury following a mixture of traumatic and vascular etiologies. Total training time was between 3.3 hours and 341.3 hours with a weighted median of 14 hours and the majority of studies carried out training in for hospitalized patients (14 studies) or in a clinic for outpatients (7 studies). Studies targeted attention or subcomponents of attention (15 studies), memory (11 studies targeted either Memory, Working Memory or Prospective memory) or simply cognition in general (7 studies). PSS CogRehab was used as training task in 4 studies while CogMed and AIXTENT were used in three studies each. Information about therapist involvement and/or blinding was only obtainable for little over half of the studies.

paper	contrasts	patients	N	age	incidence	task	target	intensity	commercial	therapist	blinded	data
Sturm et al. (1983)	NA	abi	30	44.1	299	WDG + Cognitrone	A	0.5*14=7	yes	%	%	norm
Malec et al. (1984)	NA	tbi	20	30.0	80	Target fun	Sustained A	0.4*6.5=2.7	no	no	no	event
Kerner et al. (1985)	2	tbi	24	NA	NA	Gianutsos	M	0.8*12=9	yes	%	%	norm
Wood et al. (1987)	NA	abi	20	27.8	973	Attention Retraining	Cognition	1*20=20	no	%	yes	F
Middleton et al. (1991)	2	abi	36	27.0	1096	PSS CogRehab + Sunburst	A+M / reasoning	NA*NA=32	yes	%	%	change
Sturm et al. (1991)	NA	vascular	35	50.7	103	WDG + Cognitrone	A	0.5*14=7	yes	%	%	norm
Piskopos et al. (1991)	NA	tbi	34	26.2	2032	PSS CogRehab + Gianutsos	Cognition	2*52=104	yes	no	no	norm
Gray et al. (1992)	NA	abi	31	30.0	82	(four tasks)	A	1.3*12=15.34	no	yes	yes	norm
Ruff et al. (1994)	2	%	15	26.9	NA	THINKable	A / M	2*10=20	yes	no	%	norm
Chen et al. (1997)	NA	tbi	40	28.2	345	PSS CogRehab	Cognition	1*341.2=341.3	yes	yes	no	norm
Sturm et al. (1997)	4	vascular	38	48.0	167	AIXTENT	4 A domains	1*14=14	yes	no	%	change
Shin et al. (2002)	NA	abi	27	48.2	13	PSS CogRehab	Cognition	0.5*12=6	yes	%	%	norm
Kim et al. (2003)	NA	abi	33	52.2	13	CogRehabK	A	0.5*15=7.5	no	%	%	norm
Sturm et al. (2003)	NA	abi	33	33.9	695	AIXTENT	A	1*14=14	yes	%	%	%
Sturm et al. (2004)	NA	vascular	8	56.5	399	AIXTENT Alertness	A (alertness)	0.8*14=10.5	yes	%	%	raw
Röhring et al. (2004)	NA	abi	48	53.3	775	CogPack	A	0.8*11=8.8	yes	yes	%	event
Dou et al. (2006)	NA	tbi	26	38.1	248	CAMR	M	0.8*20=15	no	no	%	norm
Man et al. (2006)	NA	abi	48	45.1	1325	Analogy problem solving	Problem solving	0.8*18=13.5	no	yes	yes	norm
Weiand et al. (2006)	NA	abi	51	41.2	NA	RehaCom	Cognition	1*15=15	yes	no	%	change
Westerberg et al. (2007)	NA	stroke	18	54.0	611	CogMed RoboMemo	WM + Executive A	0.7*25=16.7	yes	yes	no	norm
Lundqvist et al. (2010)	NA	abi	21	43.3	1411	CogMed QM	WM	0.9*25=21.9	yes	yes	%	norm
Prokopenko et al. (2012)	2	stroke	100	62.4	6	Various	A + visual-spatial	0.3*10=3.3	no	no	no	med-iqr
Yip et al. (2013)	2	abi	55	38.2	1411	VR shopping	Prospective M	0.6*11=6.9	no	no	yes	norm
Prokopenko et al. (2013)	NA	stroke	43	63.2	7	Various	A + visual-spatial	0.5*10=5	no	no	yes	med-iqr
Åkerlund et al. (2013)	NA	abi	38	52.3	844	CogMed QM	WM	0.6*25=15.6	yes	yes	no	norm
Lin et al. (2014)	NA	stroke	34	62.8	228	RehaCom	M + Executive	1*60=60	yes	yes	yes	norm
Luca et al. (2014)	NA	abi	35	32.7	137	Various	Cognition	0.8*24=18		yes	yes	med-iqr
Zuchella et al. (2014)	NA	stroke	87	67.0	14	Various	Cognition	1*16=16	yes	yes	yes	med-iqr
Lindeløv et al. (NA)	2	stroke	19	56.1	57	N-back	WM / processing speed	0.2*20=5	no	no	yes	norm
Beugeling et al. (2015)	NA	stroke	39	60.6	909	Braingymmer	A+reason+WM	0.6*57=35.6	yes	yes	yes	norm

Table 6: *some descriptives of the studies which were included in the review/meta-analysis. "N" is the number of subjects presented as [treatment group + control group = total], "patients" is the common aetiology of the sample which was coded as either TBI (Traumatic Brain Injury), stroke, vascular (cerebral haemorrhages) or the catch-all ABI (Acquired Brain Injury) which is a mixture of any of these. "Intensity" is presented as [hours per session * number of sessions = total hours]. "Therapist" is whether there was a human intervention in addition to computer-based training.*

Table 7 presents some information which allows us to evaluate each study with respect to the four criteria. While the quantitative analysis of effect sizes is the topic of the next section, this section is a logical analysis of whether the evidence (method and inferences) supports the claims made in the papers.

paper	contrast	similar	claim	p<5%	control	inference
Sturm et al. (1983)		2/5/9/0	yes	14/16=88%	TaU	main
Malec et al. (1984)		0/1/5/0	no	0/6=0%	TaU	interaction
Kerner et al. (1985)	active	2/0/0/0	yes	2/2=100%	active	interaction
Kerner et al. (1985)	passive	2/0/0/0	yes	1/2=50%	passive	interaction
Wood et al. (1987)		0/1/5/0	no	0/6=0%	TaU	interaction
Middleton et al. (1991)	M-A	2/1/0/0	no	0/3=0%	active	interaction
Middleton et al. (1991)	reasoning	1/1/1/0	no	0/3=0%	active	interaction
Sturm et al. (1991)		6/3/5/0	no	5/14=36%	TaU	interaction
Piskopos et al. (1991)		10/1/23/0	no	0/30=0%	active	interaction
Gray et al. (1992)		0/8/14/0	yes	2/22=9%	active	interaction
Ruff et al. (1994)	attention	2/2/3/0	yes	1/7=14%	active	main / post
Ruff et al. (1994)	memory	2/9/6/0	yes	3/17=18%	active	main / post
Chen et al. (1997)		8/3/10/0	no	0/4=0%	TaU	interaction
Sturm et al. (1997)	alertness	0/2/0/0	yes	1/2=50%	active	interaction
Sturm et al. (1997)	vigilance	0/2/0/0	yes	1/2=50%	active	interaction
Sturm et al. (1997)	selective	0/2/0/0	yes	0/2=0%	active	interaction
Sturm et al. (1997)	divided	0/2/0/0	yes	0/2=0%	active	interaction
Shin et al. (2002)		5/5/2/0	yes	4/12=33%	TaU	post
Kim et al. (2003)		4/0/12/0	yes	5/16=31%	TaU	post
Sturm et al. (2003)	alertness	0/2/0/0	yes	2/2=100%	active	main
Sturm et al. (2003)	vigilance	0/2/0/0	yes	1/2=50%	active	main
Sturm et al. (2003)	selective	0/2/0/0	no	0/2=0%	active	main
Sturm et al. (2003)	divided	0/1/0/0	yes	1/1=100%	active	main
Sturm et al. (2004)		0/1/0/0	yes	1/1=100%	active	nothing
Röhring et al. (2004)		0/0/0/6	yes	1/6=17%	passive	interaction
Dou et al. (2006)		9/2/6/8	yes	9/25=36%	passive	interaction
Man et al. (2006)		0/1/0/0	yes	1/1=100%	passive	main
Weiand et al. (2006)		7/0/8/0	no	1/15=7%	active	interaction
Westerberg et al. (2007)		1/2/6/0	yes	4/9=44%	passive	interaction
Lundqvist et al. (2010)		2/3/1/0	yes	6/6=100%	passive	main
Prokopenko et al. (2012)	2-5days	0/0/2/0	yes	0/2=0%	TaU	main
Prokopenko et al. (2012)	8-10days	0/0/2/0	yes	2/2=100%	TaU	post + interaction
Yip et al. (2013)	nonAI	0/3/1/0	yes	1/4=25%	passive	post
Yip et al. (2013)	AI	0/3/1/0	yes	3/4=75%	passive	post
Prokopenko et al. (2013)		1/1/3/0	yes	2/5=40%	TaU	main + post
Åkerlund et al. (2013)		1/3/1/0	yes	1/5=20%	TaU	interaction
Lin et al. (2014)		1/1/7/0	yes	6/7=86%	TaU	main
Luca et al. (2014)		0/0/0/7	yes	7/7=100%	TaU	interaction
Zuchella et al. (2014)		2/6/7/0	yes	7/16=44%	active	main + post
Lindeløv et al. (NA)	nback	0/2/3/0	no	0/5=0%	active	interaction
Lindeløv et al. (NA)	search	0/1/3/0	no	0/4=0%	active	interaction
Beugeling et al. (2015)		0/0/4/0	no	0/4=0%	active	interaction

Table 7: *assessment of the criteria with respect to the included contrasts. "similar" is the counts of SAME/SIMILAR/DISSIMILAR outcome measures.*

Representativeness and Dissimilarity: Of the 277 targeted outcome measures, 140 were scored as DISSIMILAR, 68 as SIMILAR and 69 as SAME. There was insufficient information to score 21 targeted outcomes. With respect to Representativeness, it is too comprehensive for the current purpose to evaluate the psychometric properties of each of the 277 targeted outcome measures. A crude test is counting the number of outcome measures used and see whether they pass a certain threshold, say three or four. To simultaneously satisfy the Dissimilarity criterion, we could further require that these outcome measures are dissimilar to the training. Table 7 column "similar" list the number of outcomes which are SAME/SIMILAR/DISSIMILAR and by requiring DISSIMILAR to be greater than two (as a bare minimum), we see that 21 out of 42 treatment-control contrasts fail these criteria while another 21 pass. That is, half of the contrasts fail a priori by design to sample the transfer domain sufficiently, and thereby to provide evidence that a transfer effect took place, no matter the outcome of the tests that was administered.

For example, Dou, Man, Ou, Zheng, and Tam (2006) administered 28 tests before and after treatment. These included word lists (7 outcomes), stories (2 outcomes) and digit span (1 outcome). Training also included word lists, stories and digit span. In other words, at pretest, all subjects saw these tasks for the first time. At post-test, subjects in the treatment group had done these tasks for 20 days whereas the control group had not. A statistically noticeable difference was observed for 7 of these outcome measures, but this is highly compatible with a specific-learning hypothesis. As a side note, the 10 tests sample a smaller memory domain which might be labeled "recall sequential verbal information".

Consistency: Since most authors defined "effective" as statistically noticeable ($p < 5\%$), Table 7 lists the number and proportion of outcome measures in the targeted domain which were statistically noticeable, independent of Representativeness and Dissimilarity. This percentage should be compared with the "claim" column which is whether the authors claim to have found a transfer effect ("yes") or ambiguous/null result ("no"). As a heuristic criterion, one could say that at least 75% of the outcomes in the targeted domain should be positive in order to claim that there was a consistent positive outcome. All 12 "no" contrasts have 36% positive outcomes or less which they all rightly interpret as evidence that the targeted high-level cognitive function did not improve. However, 20 out of the 30 "yes" contrasts had less than 75% positive outcomes. Strikingly, 18 of these had less than 50% positive outcome measures, yet still concluded that the cognitive function which underlies all of these had improved.

For example, the Ruff, Mahaffey, Engel, et al. (1994) memory-training group had positive effects on just three out of 17 outcome measures and it was concluded that "the training demonstrated efficacy on multiple levels" with respect to memory even though 14 memory measures did not improve.

Failing the consistency criterion does not mean that transfer to untrained stimuli did not

take place. It means that the observed pattern of positive results are too incompatible with an improvement of the targeted cognitive function to count as positive evidence.

Active control: 12 studies used active control groups, 12 used Treatment as Usual (TaU) and seven[26] used passive control groups. I have argued that closely matched active controls is required in order to claim that the *content* of the treatment made a difference in the outcomes and that it was not caused by nonspecific effects of just doing something. From this criterion, 18 studies failed a priori by design to provide convincing evidence that the computer-based task caused the improvement.

A pervasive issue in the use of control groups is how they are modelled statistically - or rather that they are not. The critical variable of interest is of course the difference between the improvement in the two groups. However, 12 publications compared significance testing on either main effects ("control group did not improve ($p > 5\%$) and treatment group improved ($p < 5\%$), therefore the treatment group improved significantly more than the control group") or post-test comparisons ("there were no difference between the groups at pretest ($p > 5\%$) and there was a difference at posttest ($p < 5\%$), therefore the treatment group improved significantly more than the control group"). Such tests are inappropriate for several reasons, which have been treated elsewhere (Nieuwenhuis, Forstmann & Wagenmakers, 2011; Redick, 2015). Briefly, you cannot support the null using frequentist p-values. Thus $p > 5\%$ is not evidence that the groups are similar. Bayesian inference is needed for such a claim (Dienes, 2011; Rouder & Morey, 2012). Second, there is no big difference between $p=4.9\%$ and $p=5.1\%$ but that would count as "significant" for these studies. For example, Prokopenko et al. (2013) found statistically noticeable p-values for both main effects and post-test comparison and concluded that the treatment group improved more than the control group. But the control groups numerical improvement was almost the same as that of the treatment group - it just scored lower at pretest. This is a likely cause of post-test difference which would have been accounted for in an analysis of interaction[27].

26 Kerner (1985) had both an active and a passive control group and is counted as both active and passive here.

27 In personal communication authors were asked to do this analysis but that request was not answered.

Summary of evidence and claims: To answer whether the evidence supports the claims, we may analyze sensitivity and specificity of the correspondence between claims and criteria satisfaction. These are compared in Figure 9 in the columns labelled "Criteria verdict" and "Paper's claim". Since any thresholds on quality are ultimately arbitrary, I present the analysis under several sets of thresholds.

As a Reference level, we may think of an intermediate level of evidence for a transfer effect as (1) three or more DISSIMILAR outcome measures, (2) an active control group where inference is based on the time×group interaction, and (3) that 75% of outcomes should show an effect. Under these criteria, not a single study has the design-result combination required to demonstrate a transfer effect (Figure 9, left panel). Six contrasts had the appropriate design to show such an effect (Piskopos, 1991; Gray, Robertson, Pentland, & Anderson, 1992; Weiand, 2006) but the results were lacking or inconsistent.

Lowering the thresholds to what may be thought of as a Minimum level of evidence, requiring just two or more DISSIMILAR outcome measures and not requiring a proper interaction analysis, does not make any contrasts pass the test (see Figure 9, middle panel). When further lowering the thresholds below the Minimum level, by accepting treatment-as-usual control groups and accepting as little as 50% of the targeted outcomes to show an effect, three contrasts (Sturm, Dahmen, Hartje, & Willmes, 1983; Lin, Tao, Gao, et al., 2014; Prokopenko, Mozheĭko, Levin, et al., 2012 8-10 days group) provided convincing evidence for a transfer effect relative to these criteria (see Figure 9, right panel). Sturm et al. (1983) was later replicated by Sturm and Willmes (1991) with a null result.

In comparison, there were 12 true negatives, i.e. null findings. Of these 4, 5, and 9 had an appropriate design to detect a transfer effect given consistent data under the Reference, Minimum and Below Minimum criteria sets respectively.

Worryingly, between 27 and 30 out of 42 contrasts were categorized as false positives. That is, they claim that there was an effect on the targeted cognitive function, but these claims were either methodologically unsubstantiated (fails Representativeness, Dissimilarity, and/or Active control) or contrary to evidence (fails Consistency). No studies were categorized as false negatives.

In summary, at the very best the literature has 100% sensitivity and 31% specificity for the detection of improvement in cross-modal cognitive functions as defined by the four criteria (see Table 8*)*. This is a very strong bias. Furthermore, false positives made up at least 858 or 69 % of the citations according to Google Scholar estimates. The vast majority - if not all - of these citations are in acceptance of the claims in the cited paper so this bias seems to extend to the scientific literature on this topic in general.

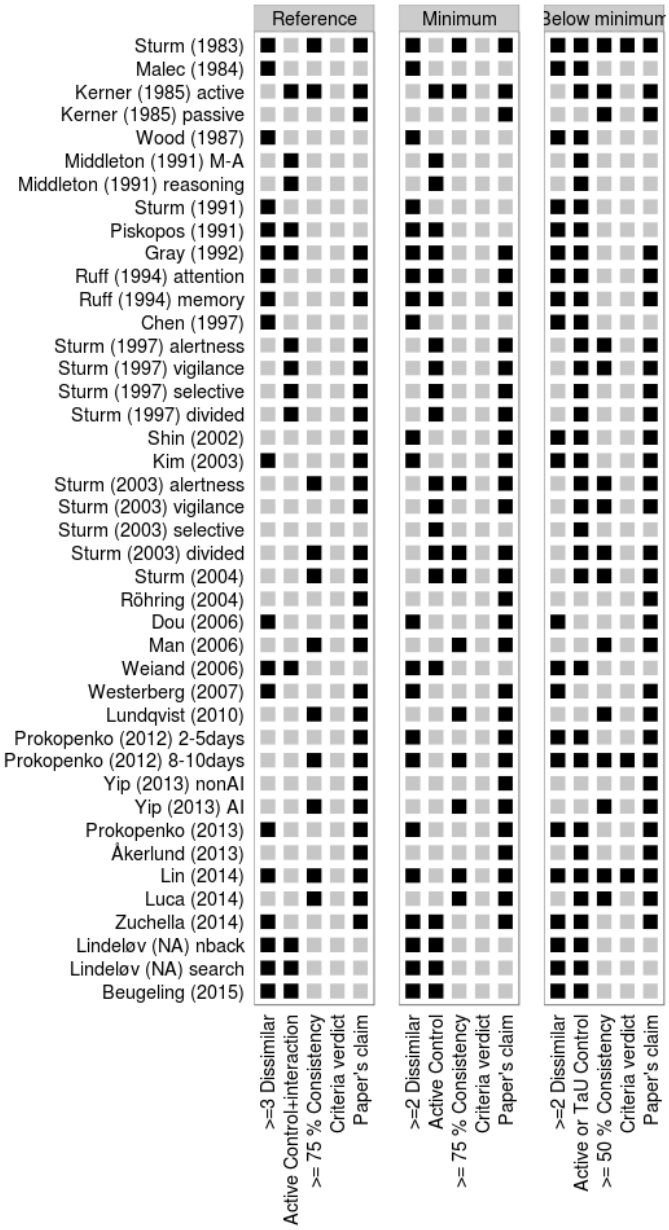
Reference
Minimum
Below minimum
Sturm (1983)
Malec (1984)
Kerner (1985) active
Kerner (1985) passive
Wood (1987)
Middleton (1991) M-A
Middleton (1991) reasoning
Sturm (1991)
Piskopos (1991)
Gray (1992)
Ruff (1994) attention
Ruff (1994) memory
Chen (1997)
Sturm (1997) alertness
Sturm (1997) vigilance
Sturm (1997) selective
Sturm (1997) divided
Shin (2002)
Kim (2003)
Sturm (2003) alertness
Sturm (2003) vigilance
Sturm (2003) selective
Sturm (2003) divided
Sturm (2004)
Röhring (2004)
Dou (2006)
Man (2006)
Weiand (2006)
Westerberg (2007)
Lundqvist (2010)
Prokopenko (2012) 2-5days
Prokopenko (2012) 8-10days
Yip (2013) nonAI
Yip (2013) AI
Prokopenko (2013)
Åkerlund (2013)
Lin (2014)
Luca (2014)
Zuchella (2014)
Lindeløv (NA) nback
Lindeløv (NA) search
Beugeling (2015)
>=3 Dissimilar
Active Control+interaction
>= 75 % Consistency
Criteria verdict
Paper's claim
>=2 Dissimilar
Active Control
>= 75 % Consistency
Criteria verdict
Paper's claim
>=2 Dissimilar
Active or TaU Control
>= 50 % Consistency
Criteria verdict
Paper's claim

Figure 9: *Criteria ratings for each experimental treatment-control contrast. The leftmost three columns in each panel is the three operationalizations presented in the main text: Dissimilarity+Representativeness, Control group, and Consistency. The rightmost two columns compare the criteria verdict (whether all criteria are satisfied) with the claim made in the publication. A black square means "yes" or "passed" while a light square means "no" or "failed". Each panel applies different sets of thresholds for these criteria. These sets are labelled in the panel title.*

"Reference" and "Minimum" criteria:

	Criteria satisfied	**Criteria not satisfied**
Effect claimed	0 true positives	30 false positives
Effect not claimed	0 false negatives	12 true negatives

"Below Minimum" criteria:

	Criteria satisfied	**Criteria not satisfied**
Effect claimed	3 true positives	27 false positives
Effect not claimed	0 false negatives	12 true negatives

Table 8: *Contingency tables with the number of contrasts categorized in each criteria x claim cell. Sensitivity is undefined for the Suggested and Minimum criteria which have a specificity of 28.6% For Below Minimum criteria, sensitivity=100% and specificity=31%.*

In this section I have argued that the claims and interpretations laid out in the papers is a poor test of a true underlying effect. This raises strong doubts about the current state of the evidence. It does not, however, inform us directly about whether transfer effects occurred in these studies or not, although the lack of consistent positive results in the transfer domain of methodologically good experiments suggest that the transfer effect is small or zero. The more direct quantification is the topic of the next section where evidence is synthesized across studies in order to quantify the actual transfer effect when discounting methodological confounds.

2.4.4 Meta-analysis: estimating the transfer effect.

Even if individual studies do not present self-contained convincing evidence for a transfer effect, an effect could still emerge when synthesizing evidence across studies. The purpose of the meta-analysis is to estimate the magnitude of the transfer effect and of the control group and similarity nuisance effects.

Statistical models

Controlled standardized mean difference: all analyses use the controlled SMD as dependent measure ($SMD = SMD_{treat} - SMD_{control}$), i.e. SMD is how much the treatment group improves when factoring out non-specific effects captured by the control group.

The unbiased population standard deviation was used as the standardizer to allow comparisons across scales and outcome measures. The population standard deviation was calculated from the unbiased pre-test standard deviations as recommended by Morris (2008) under the assumption that both treatment and control subjects were sampled from the same population (see supplementary B). This SMD is also known as the corrected Hedge's g or g*.

Norm set, full set and assumptions of normality: Authors presented data either as "norm" (20 studies; means and SD for pre- and posttest), "med-iqr" (3 studies; median and interquartile range for pre- and posttest), change scores (3 studies; mean change and SD of change) or "event" (2 studies; number of patients improved, status quo and deteriorated) (see Table 6). All analyses are carried out on a "norm" and a "full" dataset separately. The norm set comprised the 18 "norm" publications where standardized mean differences (SMD) could readily be calculated. The full set is all 26 publications, where additional assumptions had to be used to transform the 8 non-norm studies into SMD_c. Supplementary B details these transformations but briefly, these analyses assume (1) that median-interquartile represents a normal distribution, (2) that change scores follow the relatively consistent $SD_{pre}/SD_{change}=1.74$ ratio derived from known data, and (3) that event data was generated from a normal distribution. This is bad statistical practice but the "norm" studies

obviously break these assumptions as well. For example, at least half of the "norm" studies (Sturm et al., 1983; Piskopos, 1991; Sturm & Wilmes, 1991; Sturm et al., 1997; Chen, Glueckauf, & Bracy, 1997; Shin, Ko, & Kim, 2002; Kim, Ko, Seo, et al., 2003; Weiand, 2006; Yip & Man, 2013; Lin et al., 2014) report reaction time data using means and standard deviations even though a histogram or a look at the literature would reveal these to be highly right-skewed (see e.g. Ratcliff, 1979 or Whelan, 2008). The same problem usually applies to time-since-injury descriptives. Since the majority of studies potentially violate the assumptions of normality to some degree, violations of assumptions cannot be used as a strong argument against at least considering the evidence from the 8 "non-norm" studies.

The norm set should be regarded as primary evidence while the full should be regarded as confirmatory when taking all evidence into account.

Inference models: Parameter estimation was done using the MCMCglmm package (Hadfield, 2010) and inference was done using the metafor package (Viechtbauer, 2010) in the statistical software package R (version 3.0.2). Parameters were estimated using two models. (1) an intercept-only model with random effects for outcomes and studies (Riley, 2009; Ishak, Platt, Joseph, Hanley, & Caro, 2007). I call this the "naive" model because it lumps together all outcomes and designs as true reflections of an underlying cognitive construct. It is the conventional model for meta-analysis. (2) A "full" mixed effects meta-regression model with random outcomes, random studies, and covariates for control group (to quantify the Active control criterion) and difference score (to quantify Dissimilarity). Consistency is somewhat operationalized by the parameter estimates when considering all the evidence. Representativeness could not be included in the model since there is insufficient data on loading of each of the outcome measures on the cognitive constructs they target[28].

Likelihood Ratio Tests (LRT, see Barr, Levy, Scheepers, & Tily, 2013) were used to make inferences on the hypothesis that each fixed effect is zero and the strength of this evidence is quantified using bayes factors based on the the unit information prior (Wagenmakers, 2007). Briefly, a bayes factor of 5 means that the model with the parameter is 5 times more likely given the data than the model without that parameter, which has a posterior odds of ⅕=0.2. LRT tests of control group, difference scores and therapist involvement were pre-planned. Other post-hoc tests of design parameters such as blinding, treatment duration, etc. were not.

The data and analysis script is available at https://osf.io/zxgjn. I encourage you to conduct additional analyses on this dataset and to elaborate or criticise the findings reported here.

28 This analysis would entail weighting each outcome measure by a loading found through e.g. latent-variable analyses and also considering the convergence of multiple such outcomes within each study.

Results

Figure 10 and Figure 11 show forest plots of the norm set and the additional contrasts in the full set respectively. The naive estimate is plotted together with estimates broken down by level of similarity. The ranking of difference scores (same > similar > dissimilar) is evident within many of these contrasts. To be specific, the effect estimate for DISSIMILAR outcome measures was less than the the naive estimate in 15 out of 18 "norm" studies which had at least one test in another DIFF category (binomial test, p_{binom}=0.008, BF_{binom}=12.3) and 19 out of 23 in the "full" set (p_{binom}=0.003, BF_{binom}=28.6).

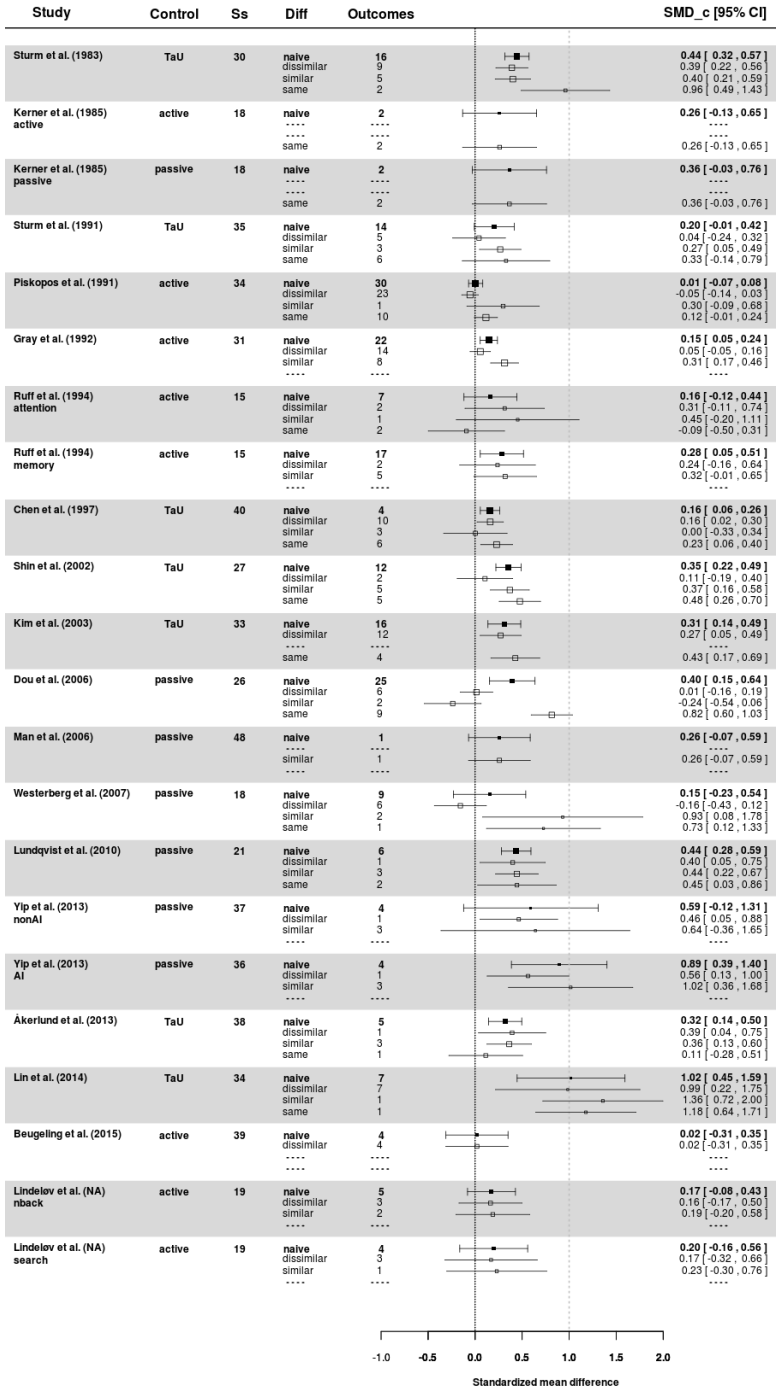
Study
Control
Ss
Diff
Outcomes
SMD_c [95% CI]
Sturm et al. (1983) TaU 30 naive 16 0.44 [0.32 , 0.57]
dissimilar 9 0.39 [0.22 , 0.56]
similar 5 0.40 [0.21 , 0.59]
same 2 0.96 [0.49 , 1.43]
Kerner et al. (1985) active 18 naive 2 0.26 [-0.13 , 0.65]
active ---- ---- ----
---- ---- ----
same 2 0.26 [-0.13 , 0.65]
Kerner et al. (1985) passive 18 naive 2 0.36 [-0.03 , 0.76]
passive ---- ---- ----
---- ---- ----
same 2 0.36 [-0.03 , 0.76]
Sturm et al. (1991) TaU 35 naive 14 0.20 [-0.01 , 0.42]
dissimilar 5 0.04 [-0.24 , 0.32]
similar 3 0.27 [0.05 , 0.49]
same 6 0.33 [-0.14 , 0.79]
Piskopos et al. (1991) active 34 naive 30 0.01 [-0.07 , 0.08]
dissimilar 23 -0.05 [-0.14 , 0.03]
similar 1 0.30 [-0.09 , 0.68]
same 10 0.12 [-0.01 , 0.24]
Gray et al. (1992) active 31 naive 22 0.15 [0.05 , 0.24]
dissimilar 14 0.05 [-0.05 , 0.16]
similar 8 0.31 [0.17 , 0.46]
---- ---- ----
Ruff et al. (1994) attention active 15 naive 7 0.16 [-0.12 , 0.44]
dissimilar 2 0.31 [-0.11 , 0.74]
similar 1 0.45 [-0.20 , 1.11]
same 2 -0.09 [-0.50 , 0.31]
Ruff et al. (1994) memory active 15 naive 17 0.28 [0.05 , 0.51]
dissimilar 2 0.24 [-0.16 , 0.64]
similar 5 0.32 [-0.01 , 0.65]
---- ---- ----
Chen et al. (1997) TaU 40 naive 4 0.16 [0.06 , 0.26]
dissimilar 10 0.16 [0.02 , 0.30]
similar 3 0.00 [-0.33 , 0.34]
same 6 0.23 [0.06 , 0.40]
Shin et al. (2002) TaU 27 naive 12 0.35 [0.22 , 0.49]
dissimilar 2 0.11 [-0.19 , 0.40]
similar 5 0.37 [0.16 , 0.58]
same 5 0.48 [0.26 , 0.70]
Kim et al. (2003) TaU 33 naive 16 0.31 [0.14 , 0.49]
dissimilar 12 0.27 [0.05 , 0.49]
---- ---- ----
same 4 0.43 [0.17 , 0.69]
Dou et al. (2006) passive 26 naive 25 0.40 [0.15 , 0.64]
dissimilar 6 0.01 [-0.16 , 0.19]
similar 2 -0.24 [-0.54 , 0.06]
same 9 0.82 [0.60 , 1.03]
Man et al. (2006) passive 48 naive 1 0.26 [-0.07 , 0.59]
---- ---- ----
similar 1 0.26 [-0.07 , 0.59]
---- ---- ----
Westerberg et al. (2007) passive 18 naive 9 0.15 [-0.23 , 0.54]
dissimilar 6 -0.16 [-0.43 , 0.12]
similar 2 0.93 [0.08 , 1.78]
same 1 0.73 [0.12 , 1.33]
Lundqvist et al. (2010) passive 21 naive 6 0.44 [0.28 , 0.59]
dissimilar 1 0.40 [0.05 , 0.75]
similar 3 0.44 [0.22 , 0.67]
same 2 0.45 [0.03 , 0.86]
Yip et al. (2013) nonAI passive 37 naive 4 0.59 [-0.12 , 1.31]
dissimilar 1 0.46 [0.05 , 0.88]
similar 3 0.64 [-0.36 , 1.65]
---- ---- ----
Yip et al. (2013) AI passive 36 naive 4 0.89 [0.39 , 1.40]
dissimilar 1 0.56 [0.13 , 1.00]
similar 3 1.02 [0.36 , 1.68]
---- ---- ----
Åkerlund et al. (2013) TaU 38 naive 5 0.32 [0.14 , 0.50]
dissimilar 1 0.39 [0.04 , 0.75]
similar 3 0.36 [0.13 , 0.60]
same 1 0.11 [-0.28 , 0.51]
Lin et al. (2014) TaU 34 naive 7 1.02 [0.45 , 1.59]
dissimilar 7 0.99 [0.22 , 1.75]
similar 1 1.36 [0.72 , 2.00]
same 1 1.18 [0.64 , 1.71]
Beugeling et al. (2015) active 39 naive 4 0.02 [-0.31 , 0.35]
dissimilar 4 0.02 [-0.31 , 0.35]
---- ---- ----
---- ---- ----
Lindelov et al. (NA) nback active 19 naive 5 0.17 [-0.08 , 0.43]
dissimilar 3 0.16 [-0.17 , 0.50]
similar 2 0.19 [-0.20 , 0.58]
---- ---- ----
Lindelov et al. (NA) search active 19 naive 4 0.20 [-0.16 , 0.56]
dissimilar 3 0.17 [-0.32 , 0.66]
similar 1 0.23 [-0.30 , 0.76]
---- ---- ----
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
Standardized mean difference

Figure 10: *Standardized mean differences and 95% confidence intervals for the "norm" studies. Each horizontal field is a treatment-control contrast. For each contrast the naive SMD_c is plotted (black and bold) followed by SMD_c broken down by similarity which reveals a same > similar > dissimilar ordering of effect sizes. Most studies have SMD_c between 0 and 1. Horizontal dashed lines indicate that there are no observations in this category. Ss = number of subjects. Outcomes is the number of outcome measures included in the analysis.*

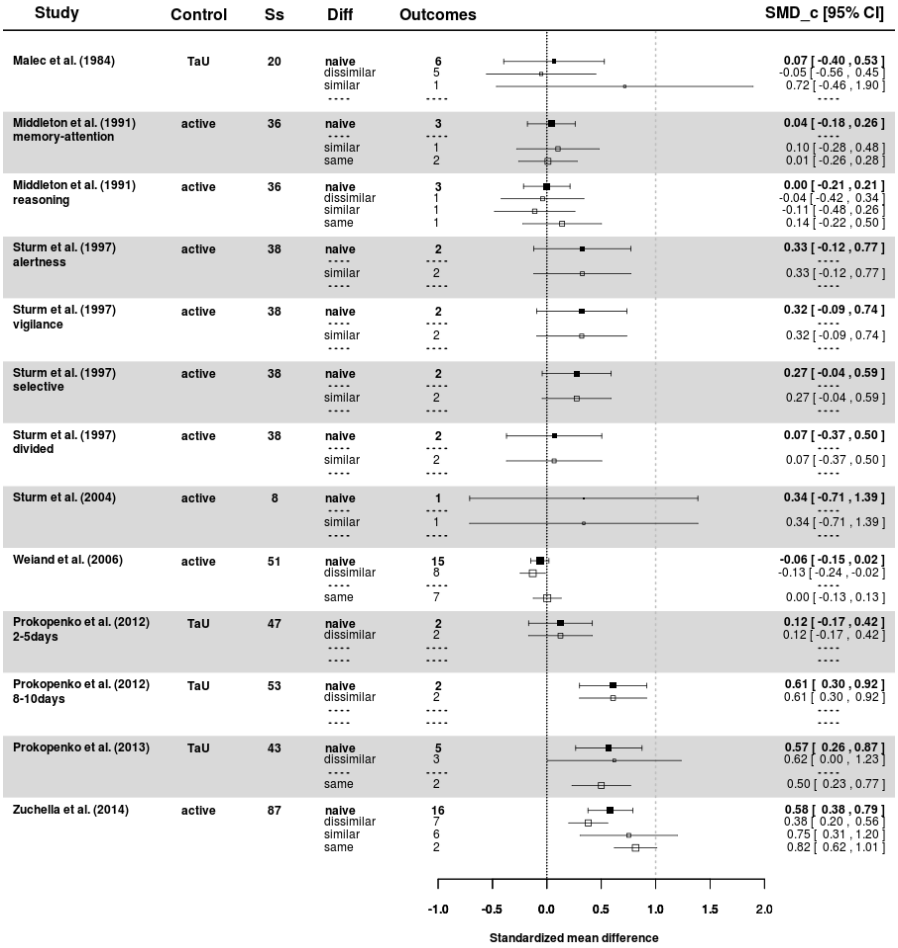

Figure 11: *Hedge's g and 95% confidence intervals for the non-norm studies. See figure text for Figure 10.*

Naive estimate: When pooling all targeted outcome measures in a single effect (naive model) there is an estimated effect of SMD=0.31 (95% CI=0.22-0.41). This effect in and of itself is conventionally interpreted as half-way between 'small' and 'medium' in magnitude (Cohen, 1992). It translates into a Number Needed to Treat of 9.57[29] (Furukawa & Leucht, 2011). In other words: for each patient who experiences a favourable outcome compared to the control group, 8.57 other patients will experience no differential effect. Still, the fact that SMD_{naive} is non-zero could be used to label computerized cognitive rehabilitation as "evidence based".

Full estimate: However, when analyzing the full model where similarity-scores and control groups are accounted for, the transfer effect is estimated to be SMD=0.06 (95% CI=-0.08 to 0.20). Adding the nonspecific effect of being in a passive control group and the specific learning of being tested on a SAME outcome measure, the gain was an order of magnitude larger (SMD=0.58, 95% CI=0.41 to 0.75). Effect estimates were almost unchanged when calculated from the full dataset. For all combinations of difference scores and control groups, see Figure 12.

Effects on SIMILAR outcome measures were larger than DISSIMILAR outcome measures (SMD=0.17, p_{LRT}=0.0001, BF=18351) but SAME was not much larger than SIMILAR in the norm dataset (SMD=0.07, p_{LRT}=0.10, BF=1.8). Effects were larger in TaU control groups than active control groups (SMD=0.19, p_{LRT}=0.03, BF=9.8) but effects in passive control groups were almost the same as TaU in the norm dataset (SMD=0.03, p_{LRT}=0.68, BF=6.2 in favor of no difference) and in the full dataset (BF=7.4 in favor of no difference).

Funnel plots for the naive and full models are available as supplementary figure 15. Heterogeneity should be expected and quantified in meta-analysis of non-identical studies (Higgins, 2008; Higgins, Thompson, & Spiegelhalter, 2009). For the full model, the variance of contrasts was $\sigma_{contrast}$=0.02 (95% CI=0.000 to 0.049) and the variance of outcome measures was σ_{test}=0.05 (95% CI=0.027 to 0.068). These variances are relatively small compared to the estimated effects (see Figure 12), so the covariate estimates are reasonably representative of the dataset as a whole (see also Figure 15).

29 Assuming a controlled event rate of 20 %

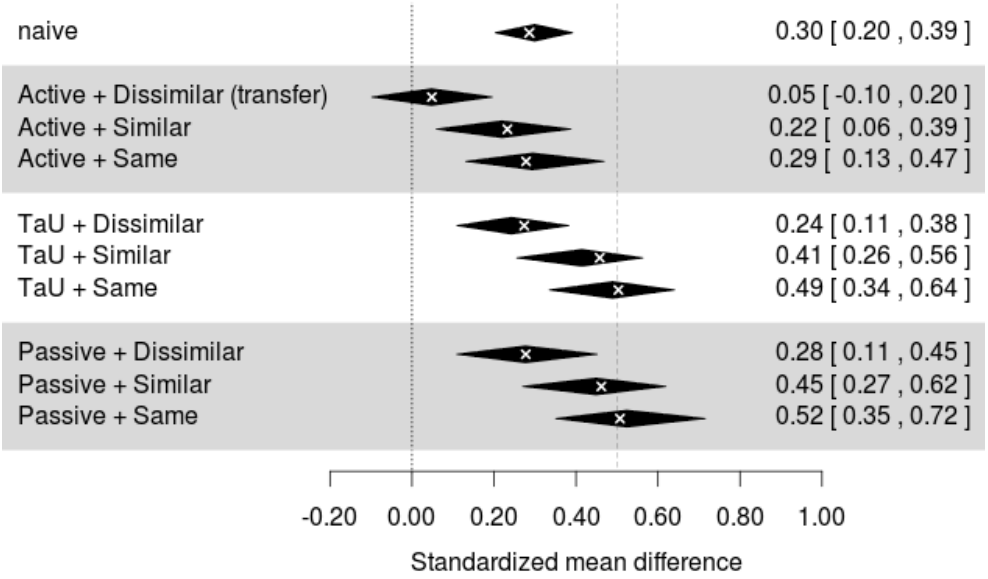

Figure 12: *predictions from a classical naive intercept-only random effects model (top) versus predictions from a mixed model with main effects of control group and outcome-training similarity. Black diamonds are means and 95% credible intervals from "norm" studies and white crosses are means based on all studies. Variance of the latter estimates are marginally smaller than the black diamonds and the estimates are similar. The crucial contrast is between the naive estimate and the transfer effect - the top two diamonds.*

Moderating factors: A large array of contrasts and models are possible using this dataset. For the sake of brevity, I will only report evidence to strengthen or weaken the finding that the transfer effect is virtually zero. Table 9 present estimates and inferences on the extent to which design- and subject-factors influence the transfer effect. Inferences are drawn from the likelihood ratio test between the full model (which is a null model in this context) and the full model plus a fixed effect for the moderating factor. All inferences were tested on the norm and full dataset, but reported for the full dataset when the two deviate qualitatively. None of the factors account for enough variability in the transfer effect to counterbalance the complexity they add to the model (all BF support the absence of moderating factors more than the presence of them), except for a negative effect of blinding as expected (Jadad et al., 1996).

Note that some p-values are less than 5% while the bayes factor weakly supports the null. This is not contradictory since p overstates the evidence against the null (Rouder, Sun, Morey, & Iverson, 2009). For example, p=5% may correspond to a probability of the null

being true between 29% and 82% (Berger & Sellke, 1987; Sellke, Bayarri, & Berger, 2001). For these reasons, bayes factors are recommended as the primary measure of relative evidence between the models.

Subject factors	stud(contr)	SMD	p_{LRT}	BF_{BIC}
Age in years (norm)	17(20)	0.003	0.45	11.0*
Age in years (all)	25(33)	0.007	0.008	2.0
Incidence in months (norm)	16(18)	-0.000	0.99	14.2*
Treatment factors	**stud(contr)**	**SMD**	**p_{LRT}**	**BF_{BIC}**
Total training hours	18(22)	-0.0006	0.17	5.6*
Number of sessions	18(22)	-0.0006	0.17	5.7*
Minutes per session	18(22)	-0.0004	0.78	14.2*
Number of tasks (norm)	13(16)	-0.001	0.06	2.1*
Number of tasks (all)	21(29)	-0.003	0.58	10.8*
Therapist, yes-no	13(16)	-0.052	0.58	10.7*
Commercial, yes-no	18(22)	-0.048	0.52	11.9*
Design factors	**stud(contr)**	**SMD**	**p_{LRT}**	**BF_{BIC}**
Blinding, yes-no	10(12)	-0.27	0.003	6.8

Table 9: *Standardized mean differences and inferences on subject characteristics and design characteristics. In general, there is little evidence that the transfer effect was influenced by these factors - most bayes factors favouring the null. Only studies which provided information about these were included in the model for each row. Stud(contr) list the number of studies and contrasts which were used for inference in each row. There are 18(22) in the norm set and 26(35) in the full set. Only inferences from the norm dataset are presented except where there is a discrepancy between the norm dataset and full dataset. BF* is the bayes factor in favor of the null. 'yes-no' indicates that the effect size is yes-no, i.e. a positive SMD is a larger effect for 'yes' than 'no'.*

2.4.5 Discussion

There is no convincing evidence that computer-based cognitive rehabilitation improves cross-modal cognitive constructs such as attention, memory and or executive function. Two independent arguments lead to this conclusion.

First, I presented four criteria (Representativeness, Dissimilarity, Consistency and Active control) as necessary premises for a convincing argument that transfer took place. They were compared to the claims made in the reviewed articles. The literature had at best a sensitivity of 100% and a specificity of 30% which is a heavy bias towards selectively interpreting positive outcomes and disregarding weaknesses and/or contrary evidence. In fact, only two out of 42 treatment-control contrasts provided convincing evidence that the computer-based training caused an improvement of the targeted cognitive function. This hardly overcomes the 12 null findings and the 28 contrasts which violated at least one criterion.

Second, controlling for non-specific effects of control groups and task-specific learning, the proper transfer effect was estimated to be between -0.09 SMD to 0.21 SMD with 95% probability. In other words, there is 2.5% probability that the effect size is larger than 0.20 SMD which corresponds to a 97.5% probability that you need to treat more than 15.7 patients for just one patient to distinguish himself/herself from the active control group. Even though the treatment durations covered a very wide range from 2.7 to 341.3 hours the duration of treatment did not influence transfer, consistent with the idea that the training task does not cause transfer. This has been found in computer-based training on schizophrenics as well (McGurk, Twamley, Sitzer, et al., 2007).

The current evidence rules out several potential "active ingredients" that one could add to the mix to stimulate transfer. There were no substantial effects of task-specific factors such as therapist involvement or commercialisation nor subject-specific factors such as age or time since incidence.

These results imply that we could invoke a law of parsimony and draw the preliminary conclusion that in general, computer-based training on patients with acquired brain injury follows the same learning principles as in healthy participants and other patient groups. I.e. that drill-and-practice interventions stimulate little transfer to high-level cognition (Perkins & Salomon, 1989; Singley & Anderson, 1989; Tournaki, 2009; McGurk et al., 2007). However, this does not preclude that such an effect could be obtained from a refined computer-based intervention and such research attempts are encouraged given the many desirable properties of a relatively flexible and automated rehabilitation method. The present findings merely implies that a priori such high-level transfer effects should not be expected given the current evidence.

Interestingly, Rohling, Faust, Beverly and Demakis (2009) found an overall effect size of SMD=0.30 in a meta-analysis of 114 studies on cognitive rehabilitation following acquired brain injury. This is exactly the same as the naive estimate in the present analysis, suggesting their estimate may be confounded by inappropriate controls and test-outcome similarities since these criteria hold for non-computerized interventions as well.

Transfer scope and strength: Specific transfer did occur as evidenced by the positive effect on SIMILAR outcomes. SIMILAR includes e.g. transfer from simple span lengths on visual digits material to simple span lengths on verbal digits, arithmetic on visually presented numbers to arithmetic on verbally presented numbers, discrimination reaction time to simple reaction time etc. That is, there does not need to be a perfect correspondence between contexts and materials for at least some transfer to take place, indicating that particular aspects of the training can indeed be extracted and brought to use in slightly deviating situations.

It is not possible to quantify the strength of transfer from the training task to SIMILAR and DISSIMILAR outcome measures since training data was rarely reported. One could use the SAME outcome measures as an index of improvements on the training task (SMD=0.29). We may heuristically quantify transfer as the ratio between gains on the training task (the transfer source) and the targeted outcome measures (the transfer domain). So 0.05 divided by 0.29 yields an estimated transfer of 17.2%. However, the SAME outcome measures are likely to underestimate the training effect since they were often administered in a different setting than the training and using a different response modality. One study (Lindeløv, submitted) explicitly analyzed and interpreted actual training data and found an improvement on the training tasks of around 0.5 to 1 SMD. This would indicate a transfer between 5% and 10% in the present case.

Revisiting the four criteria: The meta-analysis support the criteria that the SIMILAR category and the treatment-as-usual control group are impure measures of transfer effects since they deviate from the estimate of proper transfer by a magnitude of 2 or 3. This confirms the importance of satisfying the Dissimilarity and Active Control criteria. The selective reporting of positive findings on one out of several outcome measures pertaining to the same underlying cognitive constructs highlights the necessity of imposing the Consistency criterion. It is beyond the scope of this paper to quantify the effect of violating the Representativeness criterion.

I invite the reader to be skeptic about these findings and interpretations, just as I have taken a skeptical look at the literature. The strongest rebuttal would be the falsification of any of the criteria which serve as premises for most of the arguments. An easier target for criticism is the similarity ratings. I invite researchers to look at the justifications for the similarity rating of each outcome measure in the dataset (see link earlier in text). Some of these ratings were based on quite sparse information in the published literature since the

interventions are generally underspecified in the publications and had to be obtained from other sources.

In order to apply the criteria in practice, I introduced some simple operationalizations and minimum thresholds in the review section. These thresholds are ultimately arbitrary and should not be applied mindlessly, just as many other thresholds should not (alpha = 5%, BF > 3 etc.). It may be helpful for researchers to consider the four criteria but the operationalizations and thresholds are not carved in stone and should be subject to informed judgement on a case-by-case basis.

One may interpret the lack of studies passing these criteria in two ways. One interpretation is that it is very difficult to convincingly demonstrate transfer effects because they are too small or even zero. Another interpretation is that the criteria are too high. The former is supported by the meta-analysis while I have argued against the latter when I presented the criteria. For example, the Below Minimum criteria accepts treatment-as-usual which the meta-analysis showed to be confounded by other effects than the training task per se.

Limitations: The review and meta-analysis may be based on a slightly positively biased selection of the literature. Two papers (Hajek, Kates, Donnelly, & McGree, 1993; Mazer, Sofer, Korner-Bitensky, et al., 2004) were excluded solely on the basis that they assigned a modality-specific label to the training even though many of the included studies had similar unimodal training programs. Both studies were null findings. In addition, only targeted outcome measures were considered in this review so non-targeted outcome measures should per theory be subject to smaller effects. The 21 outcome measures which could not be scored on similarity had a negative effect estimate of -0.39 SMD relative to the predictions from the full model. Lastly, a simple $p < 5\%$ threshold was used to categorize an outcome as significant as did most studies. Correction for multiple comparisons should be applied to control the false positive rate in hundreds of analyzed outcome measures.

Conclusion for researchers

The findings of this review and meta-analysis lend support to the view that computer-based cognitive rehabilitation yields task-specific improvements but little transfer to untrained tasks. That is, the specificity of learning in drill-and-practice interventions is no different among computer-based interventions and brain injured individuals than in the rest of the learning literature. It is also in line with findings from recent reviews and meta-analyses on computer-based training on healthy subjects (Shipstead, Hicks, & Engle, 2012; Shipstead, Redick, & Engle, 2012; Melby-Lervåg & Hulme, 2012).

The state of the literature strongly suggests that effects of inappropriate control groups and training-outcome similarities should be expected and not confounded with proper transfer effects. To this end, I encourage considerations of Representativeness, Dissimilarity,

Consistency, and Active Control group as a supplement to the Jadad criteria.

Conclusion for clinicians

Current evidence strongly suggests that computer-based rehabilitation of brain injured patients does not improve high-level cognitive functions such as attention, memory, processing speed etc. Computer-based rehabilitation should therefore not be used for this purpose. However, this does not rule out using computer-based training of specific skills such as driving (Akinwuntan, De Weerdt, Feys, et al., 2005) or shopping (Yip & Man, 2013) using simulators which are closely matched to the environment where the patient will put such skills to use. Indeed, this meta-analysis demonstrated that even skills such as recalling digits, short messages and wordlists may be specifically trained although the amount of carry-over to everyday settings is unclear.

There was no evidence that commercial software outperformed noncommercial software. Since licenses are expensive and inflexible to handle with outpatients, I recommend using noncommercial software.

Lastly, computer-training may be used for other purposes than cognitive improvement. For example for recreation, for insight or to enhance motivation and engagement. But the evidence laid out in this paper suggest that computer-based rehabilitation is generally ineffective for rehabilitation of high-level cognition, computer-based rehabilitation.

2.4.6 Acknowledgements

Morten Overgaard and Lars Evald provided helpful comments on the manuscript.

2.4.7 Supplementary A: search and scoring

Search terms

The following search string was used on PubMed (525 results) and PsycINFO (72 results):

> *(brain-injury OR head-injury OR stroke OR concussion OR ischemic OR ischemia) AND (computerized OR computer OR computer-based OR microcomputer OR personal-computer OR pc OR tablet OR smartphone OR commodore) AND (rehabilitation OR training OR retraining OR remediation) AND (control OR controls OR passive) AND (cognitive OR cognition)*

Google Scholar were scanned for the first 500 out of 6.480 results:

> *(brain-injury OR head-injury OR stroke OR concussion OR ischemic OR ischemia) AND (computerized OR computer OR computer-based) AND (intitle:rehabilitation OR intitle:training OR intitle:retraining OR intitle:remediation) AND (cognition OR cognitive)*

Publication lists were scanned on the websites of PSS CogRehab, CogMed, and RehaCom. No studies were included from this search which were not already appraised

Study inclusion criteria

- **Intervention should predominantly be computer-based** using a monitor while physically at rest. This excludes studies with concurrent physical training, neurostimulation (tDCS, TMS), pharmacological interventions. This also excludes studies where therapist contact constituted a major part of the computer-based intervention.
- **Intervention should target higher, cross-modal cognitive functions.** This excludes skill-training programs targeting e.g. driving, computer operation skills, handling ticket automata etc. It also excludes studies targeting just visual perception or aphasia (verbal).
- **Separate control group.** Cross-over studies were included but multiple-baseline studies were not since they do not control for retest effects.
- **Sample from a fairly broad, adult, brain injured population, such as TBI, CVD, stroke or a mixture.** Studies on children or using healthy controls were excluded. Samples with specific symptoms like hemiplegia, neglect, aphasia etc. were

excluded since they are not representative of the patient group as a whole. For example, neglect patient "attention training" targeted at the neglected hemifield was not included because it might not be "attention training" for a non-neglected patient.

- **Enough data to calculate a statistic:** Single-case studies were excluded. Van Vleet et al. (2014) was not included because it was a case report with five TBIs in total: 3 in intervention group and 2 in a passive control group. Sturm et al. (2004) was included with 4 patients in each group.

Outcome inclusion criteria

- **Has to be a neuropsychological test.** Questionnaires, self-ratings, and observation measures were not included. A neuropsychological test is a test which are none of the above and which is administered by a person or a computer.
- **Has to be targeted:** An outcome measure was targeted if the authors attributed the same cognitive label to it as they did to the training task. If "cognition" was targeted, all cognitive outcome measures were scored as targeted. The vast majority of outcomes could be scored as targeted/non-targeted using these two criteria. In some publications which did not assign cognitive labels to the outcome measures, a qualitative judgement was made.
- **Has a similarity-score.**

Scoring outcomes

- **SAME:** the training task and the outcome are essentially identical. They should be identical in stimulus material, modality and in the structure of the task. For example, Lin et al. (2014) trained PSS CogRehab which includes a story recall task and used WMS Logical Memory as an outcome, which is also recall of stories of similar types and lengths. Westerberg (2007) tested subjects on a span board task but they trained on spatial span tasks as well.
- **SIMILAR:** When there is a partial overlap in stimuli and/or structure which could be explained by specific strategies that only applies to the training domain. For example, Westerberg (2007) and Åkerlund (2014) trained patients on CogMed which contains visual digit and letter span tasks while the verbal digit span tasks from the WAIS battery was used as outcome measure. Another example is Gray et al. (1992) who used PASAT as an outcome measure and trained playing an arcade game while simultaneously doing arithmetic. Here is some stimulus-overlap (numbers) and some structure overlap (arithmetic - while doing a concurrent task). But it is not the same task.
- **DISSIMILAR:** neither SAME or SIMILAR, I.e. there is no obvious low-level explanation for the observed effect on the outcome measure. For example, Sturm et

al. (1983) administered a german version of the digit-symbol coding task from WAIS. Subjects trained on various reaction time tasks where they had to detect the presence of digits and symbols, but at no time were they asked to pair them or to generate other responses than “present” vs. “not present”. Lin et al. (2014) tested subjects on a digit span tasks. Subjects trained on a forward word span task but not on digits and not in reverse or ranked order.

2.4.8 Supplementary B: Calculating effect sizes

Calculating standardized effect sizes

Effect sizes were calculated using hedge's g with the pretest SD as standardizer as recommended by Morris (2008). It is common to use a pooled SD as standardizer (the square root of the weighted mean of the unbiased variances) but this is an underestimation of the population's SD because it is assumed that there are two means and therefore the variability around each mean is lesser than if there were only one mean. Since all subjects were drawn from the same population, I think that the one-mean model better reflect the sampling intentions in these experiments. Given σ_1 and σ_2 as the pretest SDs of the treatment and control group respectively, μ as the corresponding means and n as the number of patients in each group, the population SD was calculated as:

$$SD_{population} = \sqrt{\frac{n_1(\sigma_1{}^2-1+\mu_1{}^2)+n_2(\sigma_2{}^2-1+\mu_2{}^2)}{n_1+n_2-2}} - \left(\frac{n_1\mu_1+n_2\mu_2 n_2}{n_1+n_2}\right)^2$$

… which is just the unbiased weighted mean of the sum of the squared standard deviation and the squared means minus the square of the weighted means.

The standardized effect size is the difference between pretest and posttest divided by the population SD. However, pretest and posttest are correlated because they are calculated on the same individuals. This in turns means that the variance of the estimated effect sizes are smaller than if these measures are not correlated. The correlation must be taken into account. However, the pretest-posttest correlation is not given in any of the publications.

Based on our own data from 7 different interventions, 104 brain injured patients and 126 outcome measures in this kind of studies, I calculated the pre-post correlation on each of these 126 outcomes. Correlations are essentially F-distributed so they were log-transformed to make them well represented by a mean and intervals. The correlations turned out to be remarkably consistent with a mean of 0.84 and CI from 0.82 to 0.86. They are plotted on Figure 13 along with the estimate, CI and SD.

This estimate was used in the analysis.

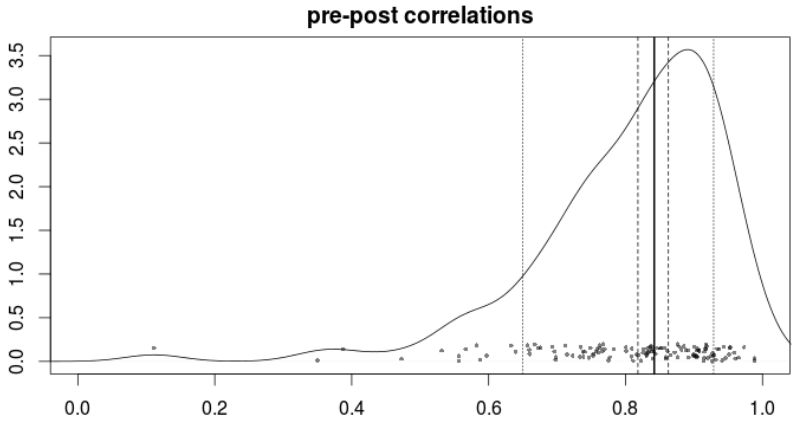

Figure 13: *Correlations between pre-test and post-test on 126 outcome measures (dots). The curved line is the empirical density of these correlation coefficients. The mean (thick black line), 95 % confidence interval (broken lines) and SDs (outer dotted lines) of the average correlation coefficient are plotted.*

Median-Interquartile data: med-IQR is often reported when the data cannot be assumed to be normal. I justify converting it to normal data for several reasons. (1) one study did not mention (non-)normality or any justification for using med-IQR instead of mean-SD (Zucchella et al., 2014). (2) two studies simply reported that data were "non-parametric" (Prokopenko, 2013) or that it failed the Shapiro-Wilk test for normality (De Luca et al., 2014). It is not clear that all outcomes were non-normal in the latter two studies. (3) data which is known to be non-normal was reported in many of the "norm" studies, e.g. non-log-transformed reaction time data. Indeed, most of these do not mention or assess normality. In summary, this conversion can be justified on the ground that the data reported using med-IQR data may not be very different from the data reported using mean-SD.

If the med-iqr is assumed to be normal, it can be transformed by setting median=mean and SD=IQR/(cdf(0.75) - cdf(0.25)) where cdf is the cumulative density function of the normal distribution.

Change-SD data: the pretest SD is needed as a standardizer in order to calculate a standardized effect size. This is, however, not given in change-SD studies, where the SD is calculated from the *change* score. I used the dataset described above to calculate the SD_{pre}/SD_{change} ratio for each of these outcomes. Since ratios typically are F-distributed, the 126 ratios were log-transformed so that it could be represented by a mean and an interval.

The ratio turned out to be relatively consistent across interventions and outcome measures with a mean of SD_{pre}=1.68*SD_{change} and CI between 1.58 and 1.79. I therefore used this estimate to convert SD_{change} to SD_{pre} and thereby calculate a standardized effect size from change-SD data.

Figure 14 shows the 126 outcomes on a log-log plot with the estimated ratio, the +/- CI and the mean +/- 1 SD. Notice how well this ratio applies across different magnitudes.

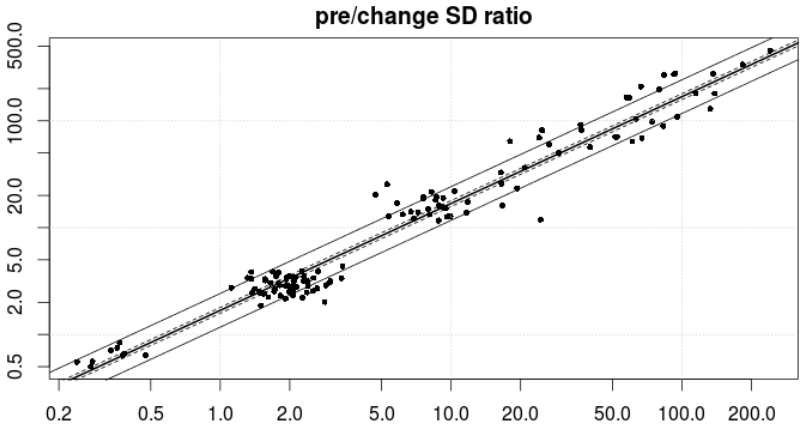

Figure 14: *log-log plot with observed data (points) and SD_{change} on x-axis, SD_{pre} on y-axis. A prediction line for the mean ratio (thick line) +/- 1 CI (broken lines) and +/- 1 SD (hard outer lines).*

Event data: Event counts were probit transformed to an estimate of SMD. This assumes that the events were discretized from a normal distribution.

Funnel plots

The outcomes are clearly heterogeneous but not biased as tested by the slope of means as a function of variances.

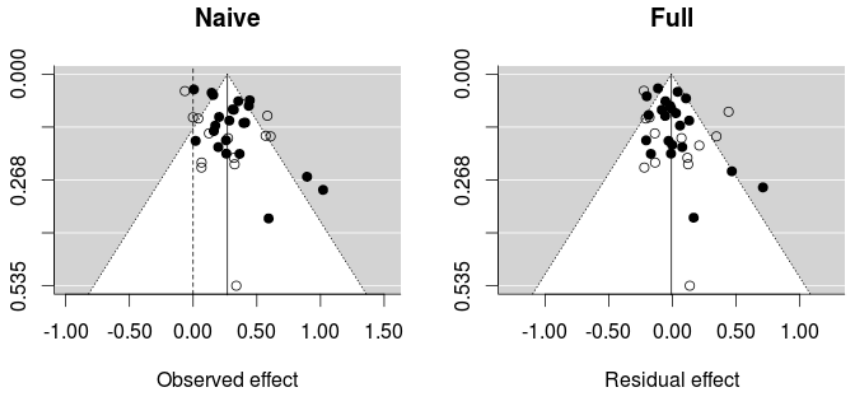

Figure 15: *Funnel plots for the naive model (left, random effects intercept) and the full model (right, random effects with intercepts for control group and similarity score). Filled circles are contrasts from the norm dataset and open circles are contrasts not in the norm dataset.*

2.5 Epilogue on computer-based cognitive training

2.5.1 The generality of the four criteria

The criteria are not limited to brain injured patients. Neither are they limited to computer-based interventions. Applying them on computer-based cognitive training of brain injured patients was simply a manageable task and a subject on which I had first-hand experience. Computer-based interventions have the advantage that they are easier to describe than therapist-administered interventions, where the human factor adds more unknowns and variability in the actual intervention. With a better description of the intervention, the scoring of similarity is more accurate and verifiable.

Being nothing more than an explicitation of the experimental method for behavioral research, the four criteria apply a to all interventions with behavioral outcomes. Sometimes the intervention obviously does not overlap with outcome measures (e.g. pharmaceuticals or exercise) and sometimes the transfer domain will be much more narrow than the ones I reviewed here (e.g. driving skills, simple math, using a Word Processor, etc.), but the same principles apply.

When there is no intervention, the Dissimilarity criterion may be dropped from the equation. This is the case in many experiments in cognitive science. Examining some classical studies in cognitive science, we see that while the Consistency and Active Control criteria are often well satisfied, the Representativeness criterion is not. The opposite is often the case in intervention research: the Consistency and Active Control criteria are often not satisfied while the Representativeness is an ideal in the sense that many different outcome measures are expected for a solid argument. In that way, basic research and intervention research may be mutually inspiring in the search for a good experimental design. To mention a few famous findings from basic cognitive science:

- Sperling (1960) is known for the discovery of iconic (or sensory) memory - an unlimited memory trace which decays largely within a second. The experiment is entirely composed of single letters and digits presented in a visual grid of various dimensions for various durations. The contrasts (Active Control) are very precise and the almost limited capacity is demonstrated under all experimental conditions (Consistency). However, the claim is that there is a short unlimited sensory trace regardless of stimuli while it was only demonstrated for visual digits and letters. It

might be that the conclusion holds only for digits and letters and does not generalize to audition, tactility, etc. The findings has since been replicated numerous times on different stimulus materials, but failing the Representativeness criterion means that the study does not hold self-contained convincing evidence that these findings generalize.

- Tulving & Thomson (1973) demonstrated encoding specificity using lists of 24 words and various cues. Again, the contrasts were well defined and the results were consistent under various cue conditions. But it was only demonstrated on lists of 24 words between 3 and 6 characters long, yet the claim generalizes to all information.
- The book chapter by Baddeley & Hitch (1974) popularized the Working Memory construct using only auditory and visual words and letters. The initial study, important as it is, did not provide self-contained evidence that the findings on the sampled material would hold for the full hypothesized domain of working memory.

All of these studies may be said to fail the Representativeness criterion at the outset, but numerous replications and extensions have indeed proven the phenomena to generalize across stimuli and contexts. This is an important point: failing anything but the Consistency criterion does not mean that the the conclusions are false. It means that the paper does present sufficient evidence to support the conclusion. Reviews and meta-analysis can transcend these limits of the individual studies.

Some classical studies do well with respect to Representativeness. For example, Miller (1956) reported a memory capacity limit of around 7 items for auditory pitch, auditory loudness, saltiness (taste), and spatial positions. Cowan (2000) sought to revise the capacity limit to a capacity limit of 4 chunks using evidence from more than 30 experiments in a wide variety of setups.

2.5.2 Is strategy learning impaired in brain injured patients?

Returning to the topic of this thesis, we observed that brain injured patients improved much less on the training tasks compared to non-injured subjects in the N-back study. Combined with the interpretation that improvements on the training tasks were highly specific (see N-back epilogue), I argued that this impairment indicates an impairment of the ability to form simple and effective strategies. Space limited the presentation of the full argument for this claim. Here I will expand on the argument. I will begin by applying Morgan's canon: is there a simpler explanation for the observed results? Such simple explanations could be experimental confounds:

Age: The most obvious confounding difference between the patient group and the healthy group is the age differences of 27 years. However, the review found no effect of age on SAME outcome measures, which is our closest estimate of the training effect in the reviewed studies in the face of no published training data. In fact, the full dataset showed a weak trend in the opposite direction: that improvement on SAME outcome measures increases with age (p_{LRT}=0.008, BF=2). The effect was only 0.007 SMD per year which amounts to a predicted SMD=0.19 difference in favor of the brain injured group in the N-back study[30]. This is clearly a long way from the observed difference of SMD=5.6 and SMD=2.5 in favor of the non-impaired group. This weak age advantage was also found in healthy subjects by Schmiedek et al. (2010) who subjected 101 young (25.4 years) and 103 old (71 years) subjects to a 100-hour intervention. They found an average increase on the training tasks of 0.81 SMD for the younger subjects and 0.98 SMD for the older subjects. Dahlin et al. (2008) observed no difference between 15 younger and 13 older adults on four training tasks and a small advantage for the younger group on one training task. Based on these data, I find it unlikely that age is a major confounder of the difference in improvement on the training task.

Training location: In the N-back study, patients trained at the hospital while the non-impaired subjects trained at home. Had they shifted training location, would the results have shifted likewise? Although it is not clear how that should influence learning during training, it might nonetheless be a factor which should not be attributed to the condition of having brain injury per se.

We can use the review data to get an estimate of the influence of training location. It was possible to obtain information about training location for 24 studies and 33 contrasts. There were four different locations: home (as the non-impaired group in our N-back study), hospital (as the brain injury group), lab (patients living at home but training at the experimental facilities) and home+lab (a combination of the latter two). Testing a model with an interaction term for SAME □ location, the model without location was clearly preferred (BF=265 in favor of the null with a unit information prior on covariates). The estimated effect sizes of different locations was at most 0.30 SMD. The bayes factor enforces parsimony so the sheer number of location-variables could cause a preference for the null even if they individually explain a bit of the residual variance. Reducing the five location to a single numeric predictor with the hypothesis that SAME □ location follow a ranked order (effect size for home > home + hospital > lab > hospital) was not preferred either (SMD=0.06 per category, BF=6.3 in favor of the null). The training location hardly explains the observed differences in training gains.

30 (56 years - 29 years) × 0.007 SMD/year = 0.19 SMD

Treatment as usual: Our N-back patients participated in rehabilitation efforts and were assisted with activities of daily living besides the computer training. Presumably, the healthy participants did not. Would we have observed the opposite training improvement pattern if the two groups had exchanged daily activities?

The review dataset allows for a test to see if this holds in general. The treatment-as-usual hypothesis can be operationalized as the interaction between TaU and SAME outcomes, i.e. whether the presence/absence of treatment-as-usual makes a difference for the change in SAME outcomes. Treatment □ SAME does not explain any variance over and beyond main effects of control group and similarity (BF=7.1 and 9.1 in favor of the null for the norm and full dataset respectively). So this is not a confound in other studies.

Likewise, there are indications that it does not in our study. (1) All the healthy participants were students or working, thus certainly active and learning most of the day. If they can do this and improve on the training task, it is not clear that replacing their daily activities with physiotherapy, exercise, group discussions, etc. would prevent them from improving. (2) The patients had a wide range of activity levels, from 1-3 hours daily to a full schedule with visits in the evening. Still, no patient reached more than N=6 in the N-back task (compare to N=14 in the non-impaired group, after removing the outliers that reached level 27 and 46) and N=50 in the Visual Search task (compare to N=121 in the non-impaired group). (3) Several of the patients requested more training, indicating that they did not perceive the treatment-as-usual as a barrier.

What is the cause of the difference? Being unable to attribute the impaired task-specific learning to confounders, I conclude, as I did in the paper, that the cause most likely has to do with the brain injury, i.e. a property of the subject. The present evidence does not permit taking this conclusion further. However, concluding that learning on computer-based training tasks is impaired in brain injured subjects relative to healthy participants is a relatively big step if it replicates. It did hold across two different training tasks in the N-back study, so this may count as a semi-replication. But given the small sample size, this single study merely puts the idea on the table and further studies will substantiate or discredit this idea.

The review was carried out after the N-back study was finished. Looking back, it is interesting to see how the N-back study fared with respect to the four criteria. The overall scorings are shown in Table 6, Table 7, and Figure 12 in the review paper. The N-back study was classified as a True Negative in Table 8. A more qualitative analysis is the topic of the next few sections:

2.5.3 Representativeness for working memory

First a warning: the length of this and the next section on the representativeness of just the tests used in the N-back study is a testament to the amount of work needed to be done to get just an approximate grasp on the representativeness of the test batteries for all 30 studies in the review. This is the reason that such an analysis was omitted.

The domain of working memory was sampled using five different tests:

- WAIS-IV Digit span forward, backward, and ordered - age corrected.
- WAIS-IV Mental arithmetic - age corrected.
- WAIS-IV (optional) Letter-number sequencing, age corrected.
- Operation Span with partial credit unit scoring (Conway, 2005).
- Raven's advanced progressive matrices (RAPM),

C.f. Morgan's canon, we should use the lowest possible model to account for the constructs in play in this test battery. The lowest-level heading that represents these five tests could be "temporary processing and storage of visual and verbal information". This is relatively close to classical definitions of working memory. Were it not for RAPM, the heading could have been more specific: "temporary sequencing/arithmetic and storage of visual and verbal digits/letters". These tests do not sample the full domain of working memory since there is no assessment of tactility, olfactory working memory (Dade et al., 2001), tactility etc. A few comments on the properties of each test with respect to a hypothesized underlying working memory construct are in order:

WAIS working memory subtests: A first quality of the WAIS tests is that there is a danish normative reference material to make scores comparable and to correct for age influences. Second, the working memory and processing speed indices has been shown to have good internal consistency (> 0.85) and test-retest reliability (> 0.75; Iverson, 2001). Numerous studies have tested the factor structure of WAIS and generally found that the Working Memory Index indeed explains variance not captured by other tasks or indices (e.g. Canivez & Watkins, 2010; Benson et al., 2010; Holdnack et al., 2011). There are nuances to this story. The high loadings of the Working Memory factor on a general intelligence factor ($r > 0.90$ in all studies) means that when g is accounted for, working memory accounts for little of the remaining variance ($r < 0.25$). This lead Canivez & Watkins (2010) to suggest that g should be the primary level of inference.

The mental arithmetic subtest generally has a smaller load on the working memory factor than the other subtests. It could be that the shared variance between these other subtests, which are all composed of sequences of letters/digits, can be attributed to shared low-level properties rather than a high-level working memory construct. If this is the case, the lower

loading of Mental Arithmetic (down to r=0.57) could be most representative of the proportion that is shared through a common high-level factor.

Operation Span is one of three common complex span tasks: reading span, symmetry span, and operation span. Of these, Operation Span has shown to share most variance with a common underlying construct, although only marginally so (r=0.76 vs, 0.72 and 0.73; Conway et al., 2002; Engle et al., 1999). This complex span-derived working memory construct could again be said to be confounded by shared low-level commonalities because the three complex span tasks are structurally identical, just using different stimuli. Here also, the working memory factor is almost identical to a fluid intelligence construct (r=0.98).

Operation Span was initially presented by Daneman & Carpenter (1980) and was later shown also to have good psychometric properties in a computerized version (2005). It has become the "Gold standard" operationalization of working memory in healthy populations but it is rarely applied in patient populations.

RAPM is often called the "gold standard" operationalization of fluid intelligence given it's very high loading on such a latent construct with r>0.90. The reason for including it in a test battery for working memory is that Engle et al. (1999)'s and Conway et al. (2002) entertained the idea that Gf might be the central executive in the working memory models by Cowan (1988) and Baddeley (1974, 2007). To be clear, they speculate about the specific claim that Gf=Central executive and not the more general claim Gf=working memory. Correlations for the latter relationship is far lower than suggested by unity (r=0.5 to 0.6; Ackerman, Bier & Boyle, 2005). Thus while RAPM can be said to be in the working memory domain, it shares most of its variance with the executive part of working memory in particular.

Representativeness verdict: In hindsight, I would say that this "working memory test battery" is representative enough to at least stimulate curiosity about an effect on the level of working memory if a consistent effect was found. Given the low-level similarities of the WAIS span-subtests and the tasks used to infer the loading of Operation Span on a working memory construct, I would not say that this test battery is suitable for definitive claims about high-level effects. However, it should be noted that the test batteries of many studies would fail at this level of scrutiny.

2.5.4 Representativeness for processing speed

While the Visual Search task was initially thought of as an active control, it does have some of the characteristics of a processing speed task. N □ N stimuli has to be scanned within less than two seconds while the N-back task only requires consideration of N stimuli. To test for the hypothesis that visual search training improves processing speed, three tests were added to the battery:

- WAIS-IV Symbol-digit coding - age corrected.
- WAIS-IV Symbol search - age corrected.
- Computerized stroop color-word interference: stroop effect.
- Computerized stroop color-word interference: reaction time.

The lowest level heading that subsumes these outcome measures could be "speed of judgements on simple symbols and colored words". All tests are visual. Lacking e.g. auditory tests and visual images, this battery is not as representative as one may require to establish claims about improved processing speed, even though it passes the criterion of "at least 3 dissimilar tests". Still, it is sufficient to falsify a claim about transfer. If these tests fail, one would need 12 other positive outcomes to make the compound result passe the Consistency >= 75% level.

WAIS processing speed subtests: As with the WAIS working memory index, the processing speed index has also been shown to explain unique variance. The two subtests share many low-level features which may account for at least some of this shared variance. For example, both involve simple visual symbols and both involve a great deal of horizontal left-to-right scanning and motor response. Interestingly, processing speed shares almost no variance with fluid intelligence ($r < 0.11$) and little with working memory ($r < 0.29$; Conway et al., 2002), making this index relatively independent of the working memory index (Canivez & Watkins, 2010). This is a good thing with respect to testing differential effects in the domain of processing speed and working memory respectively.

Stroop: This is seldomly used as a processing speed measure but it is readily interpretable as such given the central role of inhibition in race-based cognitive theories (e.g. Norman & Shallice, 1986; Cowan, 1988) at which schemas or chunks are selected by effectively inhibiting competitors.

Absolute stroop reaction time can be thought of as a continuous-performance task. In earlier experiments I have found the stroop effect and absolute reaction times to be independent and therefore interpretable as different outcome measures. This did not hold for the N-back studies where I found a linear relationship, meaning that to some extent the

two indices represent the same ability, just measured at different scales (see Figure 16). Since they arise from the same test, they share maximal low-level task similarity and are thus not readily interpretable as separate measures of a high-level construct.

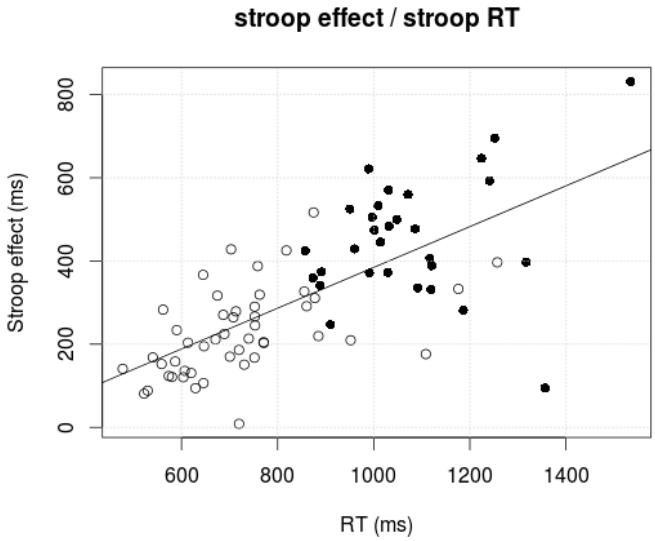

Figure 16: *Scatter plot of the stroop effect as a function of absolute reaction time as measured in the N-back studies. Open circles are healthy subjects. Filled circles are patients. A common linear regression line is superposed. It is clear that the the stroop effect and the reaction time covaries and that patients were slower than healthy subjects.*

Representativeness verdict: In hindsight, I would say that the processing speed test battery fails to represent a broader domain of processing speed. Given the low-level similarities between the WAIS subtests and the correlation of stroop measures, it does not constitute four low-level-independent tasks. Still, a consistent positive finding on these tasks could stimulate further research to assess the generality of such a finding.

2.5.5 Dissimilarity

In the review process, I explicitly wrote a motivation for the similarity-score for each outcome. They can be seen in the dataset to the review in the column “diff.text”. I copy-paste them in here for completeness. For the N-back training there were three dissimilar and two similar outcomes:

- Digit span was scored as SIMILAR. N-back on digits was trained. Though as passive recognition and seldomly higher N than 4.
- Arithmetic was scored as DISSIMILAR. No arithmetic in training. Some numbers, but no counts of objects as in the mental arithmetic test.
- Letter-number sequencing was scored as DISSIMILAR. No sequencing in training. In N-back, things just have to be recalled in the presented order.
- Operation Span was scored as SIMILAR. Some resemblance to the letter-condition during training.
- RAPM was scored as DISSIMILAR. There was a spatial N-back condition, but no hidden rules or pattern recognition in simultaneously presented stimuli.

and for the Visual Search training there were three dissimilar and one similar outcome:

- Symbol-digit coding was scored as SIMILAR. Resembles the training task but differs in that there are two targets and the layout is different and you have to indicate location rather than mere presence.
- Symbol search was scored as DISSIMILAR. Training did not involve numbers, nor the drawing of shapes.
- Stroop effect was scored as DISSIMILAR. No color and no words in training.
- Stroop reaction time was scored as DISSIMILAR. No color and no words in training.

Taken together, there are three dissimilar outcome measures for each intervention. This just barely meets the “Reference” threshold level for the Dissimilarity criterion in the review paper. Note that I included both Stroop outcome measures in spite of the considerations of representativeness. I did this because it was in the supplementary to the N-back paper (I should not treat my own paper in a special way in the review) and since many studies had such multiple indices from the same task.

2.5.6 Active control

The active control group should ideally isolate the independent variable which is hypothesized to cause cognitive improvement, which we observe as improvement on the outcome measures. In the N-back study, this independent variable can be described as "the retention and updating of sequential visual and auditory-verbal information" since the visual search task had neither retention, updating or sequential information. For a perfect contrast, the visual search tasks should have had an equal number of different auditory and visual implementations with the same characteristics as the N-back stimuli. Still, the N-back study had the most closely matched control group of all studies included in the review. The second place is shared by all studies using computer training as active controls (Sturm et al., 1997, 2003, 2004; Beugeling, 2015; Ruff, 1994; Gray, 1992; Middleton, 1991; Kerner, 1985). In all of these studies, the training tasks diverged on many other parameters.

2.5.7 Consistency

There was not a positive effect on any outcome measure. Since Consistency is about consistent positive outcomes, the N-back study fails this criterion.

The total verdict of applying the "Reference" thresholds to the N-back study is that it is methodologically capable of detecting a transfer effect if there was one, but the data indicated that there was not. Therefore, we interpreted the results as a null finding with respect to transfer.

3 Suggestion-based cognitive rehabilitation

3.1 Prologue: Hypnosis and brain injury

We conducted a study where we used hypnotic suggestion to improve working memory in brain injured patients. The study itself was conceived and carried out before the studies on computer-training. It is in review in Current Biology as of writing this text. The Current Biology report format does not allow for much background information, so I will use the following sections to present the background on which this experiment was conceived.

3.1.1 Meditation and the improvement of attention

The initial idea for this study was the result of the observation that attention-related problems were frequent in (but not limited to) brain injured patients (Cumming, Marshall, & Lazar 2013; Carroll et al., 2003, Engberg & Teasdale, 2004; Fox, Lees-Haley, Earnest, & Dolezal-Wood, 1995) and that meditation was supposedly a method to improve attention directly in healthy subjects (Moore & Malinowski, 2009; Chiesa et al., 2011). Connecting the dots, I anticipated that meditation may be a good intervention for these patients. However, I soon discovered that such attempts had been carried out with consistently[31] negative results with respect to high-level cognition, such as attention and memory (McMillan et al., 2002; Bédard, 2003; Bédard, 2005; See also two recent papers with similar findings: Johansson et al., 2012. Azulay et al., 2013) although positive effects on other aspects were observed in some of these studies. Sedlmeier et al., (2012) carried out a meta-analysis on 163 meditation studies (the hitherto largest meta-analysis on the topic) and found small-to-medium sized effect on attention (r=0.28), memory (r=0.21) and intelligence (r=0.20) in healthy subjects. To put these effect sizes in a clinical perspective, these effect sizes correspond to Number Needed to Treat[32] of 5.1, 7.1 and 7.6 respectively, so less than 20% of subjects would show a differential effect compared to a control group.

3.1.2 Hypnotic suggestion and the manipulation of cognition

Exposed to the meditation literature, it was not long before I encountered the literature from the recent upsurge in basic science research on hypnosis (Oakley & Halligan, 2009, 2013). fMRI studies on hypnosis showed that highly hypnotizable subjects were able to inhibit

31 Bédard (2002) was a pilot study with positive results but a larger replication by Bédard (2005) failed to reproduce the results.

32 Assuming the default controlled event rate of 20%. If CER=50%, NNT=10 and if CER=50%, NNT=7.1.

color vision (Kosslyn et al., 2000), to induce pain in the absence of stimulation (Derbyshire et al., 2004), and to inhibit pain in the presence of stimulation (Vanhaudenhuyse et al., 2009). These studies had not only convincing behavioral results but also perfect physiological results where real perception and suggested perception were indistinguishable. Similar results were obtained in behavioral studies where the stroop effect was virtually eliminated (Raz et al., 2002, see review in Lifshitz et al., 2013 for a list of more than 10 successful replications), and the McGurk effect was reduced (Lifshitz et al., 2013). Taken together, this literature implies that hypnotic suggestion can de-automatize what was previously thought to be automatic cognitive processes to a hitherto unprecedented extent (Kihlstrom, 2011).

This literature gives the clear impression that hypnotic suggestion is a powerful means of altering human information processing. However, the literature almost exclusively targets low-level cognition and perception, and I searched in vain for studies targeting the phenomena of interest for the present purpose, namely high-level cognition such as attention, working memory, executive function etc. An exception is perhaps the failed attempts at enhancing recollection (see below). We set out to fill this gap with an experiment. Before proceeding to our experiment, a few comments on the scope and limits of hypnotic suggestion are in order:

Naturally, there are limits to what can be achieved using hypnotic suggestion compared to waking suggestion. With respect to the findings above, the induction of color blindness findings of Kosslyn et al. (2000) was not reproduced when changing the suggestion slightly (Kirsch et al., 2008; Mazzoni et al., 2009). Although the stroop effect finding is replicable, negative priming between consecutive trials was unaffected by the suggestion, indicating that subjects were not wholly unable to perceive the words presented in colored ink (Raz & Campbell, 2011). There is at least uncertainty in whether recall can be genuinely enhanced or whether apparently improved memory can be attributed to changed criteria (Dywan & Bowers, 1983; Klatzky, 1985). More generally, Ted Barber famously demonstrated that a number of hypnotic behaviors could be a result of experimental methods differing in subtle ways between hypnotic conditions and non-hypnotic conditions and he discounted the idea that hypnosis was a distinctive mental state (Barber & Gordon, 1965). Similarly, mentioning the word “hypnosis” increases suggestibility much more than if substituted with “relaxation” (Gandhi & Oakley, 2005), indicating that prior expectations about hypnotic experience and behavior is a major factor in hypnotic responsiveness.

Several theories have been put forth to systematise positive and negative findings. They will not be discussed here since our experiment was not designed with any of these theories in mind. I will briefly speculate about how our findings might inform theory in the epilogue to the paper.

For the present purpose, I will try to throw off as much theoretical weight as possible and

adopt a rather simple procedural account of what hypnosis is[33]: Suggestion is guidance of thought, usually delivered through language (White, 1941). The utterances "everything will be all right", "look, there's a bird over there" and "hand me the salt" are all suggestions, just as is "you are now in a deep hypnotic state" and "your hand is numb now" are suggestions (Hilgard, 1973). Suggestion, whether hypnotic or not, can bring about profound changes in behavior that are relatively consistent with the suggestion given (Kirsch & Braffman, 2001). The induction and termination of hypnosis is nothing but a series of suggestions, usually about relaxation and heightened suggestibility (Kihlstrom, 2008). The word "hypnosis" as in "hypnotic suggestion" and "hypnotic state" denotes a particular family of suggestion, just as "meditation" and "psychotherapy" denotes other families of suggestion.

Much is left unsaid about the hypnotic state and the mechanisms by which suggestion operate, but this short introduction will suffice for the claims I make in the paper.

3.1.3 Suggestibility of brain injured patients

Can brain injured patients be hypnotized at all? The answer seems to be yes. Laidlaw (1993) tested 21 concussed outpatients on a Harvard Group Scale of Hypnotic Susceptibility form A and found that they were no different than 31 healthy controls with respect to suggestibility. Likewise, Kihlstrom (2013) tested 15 patients on the Arizona Motor Scale of Hypnotizability (AMSH) and found them to be like healthy controls, regardless of site of lesion[34]. There are no danish norms for Stanford Hypnotic Suggestibility Scale, form C, but our 49 patients in the treatment groups scored higher than normative means from 8 out of 10 other nationalities (see Table 10. Our sample mean is 7.3 (median 8, see 17). With an average age of 46.0 years, our sample was much older than the oldest normative sample (Taiwanese, 33.6 years) but age differences are unlikely to affect the results considerably given that Piccione et al. (1989) found no time-trend in suggestibility over a 25-year period. Taken together, the evidence support that patients with mild to moderate brain injury can be hypnotized and that they may even be more suggestible than the healthy population on average.

33 Barnier & Nash (2008) distinguished between hypnosis-as-procedure and hypnosis-as-product as two different ways of accounting for what hypnosis is. The former describes the procedures and the observed behavior whereas the latter describes this in mental terms on the part of the hypnotized subject. The former often assumes the latter without testing it.

34 Kihlstrom used this sample to falsify the hypothesis that hypnosis represents a shift to right-hemisphere dominant processing.

Reference	Nationality	Age	N	Mean (SD)
Naäring et al. (2001)	Dutch	21.5	135	4.31 (2.6)
Hilgard et al. (1965)	American		307	5.19 (3.1)
Bongartz et al. (2000)	German	22.4	174	5.53 (2.8)
Lichtenberg et al. (2009)	Israeli-Hebrew Israeli-English	around 30	169 38	5.62 (2.4) 5.68 (3.3)
González et al. (1996)	Spanish	19.5	115	5.78 (3.2)
Pascalis et al. (2000)	Italian	19-29	356	6.81 (around 3)
Roark et al. (2012)	Taiwanese	33.6	322	6.87 (2.4)
Lindeløv et al.	Danish	46.0	48	7.3 (2.7)
Sánchez-Armáss et al. (2005)	Mexican	25.5	513	7.56 (2.3)
Allen (2008)	American-Indian	31.6	40	7.75 (2.9)

Table 10: *SHSS:C normative results for healthy subjects of different nationalities, ordered by mean score. Note that SHSS:C scores are not normally distributed (see Figure 17) so the means and standard deviations are not truly representative of the population.*

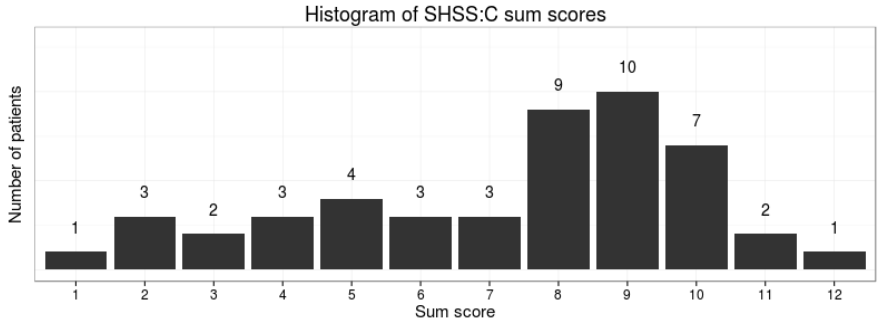

Figure 17: *The distribution of SHSS sum scores at baseline in our sample of 49 patients. The mean is 7.3 and the median is 8. This non-normal distribution is similar to the published histograms from the normative studies in Table 10.*

3.1.4 Pilot study

We ran a pilot study in the fall of 2009 and spring of 2010 including four patients with acquired brain injury. The methods and results of this pilot is described in the paper's methods and supplementary. Briefly, two patients were given suggestion which targeted working memory related experiences explicitly. Among other techniques, we used used age regression to the pre-injury state, visualizations of recovery in the present and post-hypnotic suggestion about successful experiences. A working memory index score was enhanced by 1.4 SMD (using the healthy population standard deviation as standardizer) following 8 weekly 1-hour sessions of this intervention (see figure 24, subject 1 and 2). This effect was sustained after 2 months break (timepoint 2 and 3). Two other patients (subject 3 and 4) were assigned to a control condition consisting of hypnotic induction → listening to music → termination. They remained at status quo improvement (0.07 SMD). When crossed over to the targeted-suggestion condition, they improved 1.3 SMD, similarly to the first two patients.

To put these effect sizes in perspective, they correspond to a Number Needed to Treat of 2.0. A power calculation for an interaction effect in a repeated measures ANOVA estimated the appropriate sample size to be 6 subjects in total for 95.8% power to detect an interaction effect at an alpha level of 1%. This convinced us that this project was worth pursuing. We hypothesized that the improvement was caused by the content of suggestion rather than the hypnotic state or suggestion per se. We made the appropriate corrections in the script and design to isolate that variable.

The experiment, which is reported in the paper, simply establishes the strength of the relationship between the targetedness of suggestion and the effects on outcome measures without theorizing about the cause of such an effect. Given the lack of prior knowledge on this topic, that step is big enough in and of itself so more detailed research questions about the psychologically mediating mechanisms were saved for later studies.

3.2 Hypnotic suggestion can normalize working memory performance in chronic brain injured patients.

Working memory is the mental capacity to perform short-term storage and manipulation of information in the face of concurrent processing (Baddeley, 2007, Cowan, 1988). The importance of working memory in everyday functioning (Alloway & Alloway, 2010) has given rise to a broad interest in persistent enhancement of WM in healthy subjects and in patients. Here we show that hypnotic suggestion can effectively restore otherwise impaired working memory performance in patients with acquired brain injury. Following four sessions, patients had a standardized mean improvement of 0.68 and 0.57 more than an active control group and around 1.1 more than retest. This effect was maintained after a 6.7 week break. The active control group was crossed over to the working memory suggestion and showed superior improvement. Both groups reached a performance level at- or above-the healthy population mean with effect sizes between 1.55 and 2.03 standard deviations more than a passive control group. We conclude that there may be a residual capacity for information processing in patients with mild brain injury and that hypnotic suggestion may be an effective method to improve working memory performance.

3.2.1 Introduction

Working memory is a central cognitive function that enables short-term storage and manipulation of information in the face of concurrent processing (Baddeley, 2007, Cowan, 1988). The importance of working memory in everyday functioning (Alloway & Alloway, 2010) has given rise to a broad interest in persistent enhancement of WM in healthy subjects and in patients.

Biological approaches include pharmaceuticals Repantis, Schlattman, Laisney, & Heuser (2007), physical exercise Smith et al. (2010), neurostimulation (Thut & Pascual-Leone, 2010) and nutrition Balk et al. (2007). However, these interventions do not target any particular cognitive domain and effect sizes on higher cognitive functions are in the zero-to-moderate range for repeated administrations.

Behavioral approaches targets specific mental functions by loading them bottom-up. Bottom-up refers to processing elicited by incoming sensory stimuli. This approach usually uses computerized training or other repetitive tasks to do "work out" of working memory by repeatedly straining it bottom-up through difficult tasks. However, effects tend to be

specific to the trained task, i.e. not generalizing to dissimilar tasks (Melby-Lervåg & Hulme, 2012; Shipstead, Hicks, & Engle, 2012; Shipstead, Redick, & Engle, 2012).

In this study, we pursue an alternative and novel approach: top-down enhancement of working memory capacity. Top-down processing refers to processing elicited by previous experience, expectation, context and motivation. Suggestions are probably the most direct behavioral instrument to experimentally control top-down predictions and hypnotic suggestions seem to be particularly powerful in this respect (Kirsch & Braffman, 2001). Examples include unprecedented control over color perception Kosslyn et al., (2000), the McGurk effect (Lifshitz, Aubert Bonn, Fischer, Kashem, & Raz, 2013), the Stroop effect (Raz, Shapiro, Fan, & Posner, 2002) and pain perception (Derbyshire, Whalley, Stenger, & Oakley, 2004; Vanhaudenhuyse et al., 2009). However, these were all obtained through suggestions about altered perception. The current study uses suggestions about working memory in order to investigate to what extent higher cognition can be enhanced using hypnotic suggestions.

We recruited 52 subjects with acquired brain injury in the chronic phase of which 49 completed (see descriptives in Table 11). They were randomly assigned either to receive targeted or non-targeted suggestions. The targeted procedure consisted of suggestions about enhancing WM functions. The theme of the targeted suggestions was the instantiation of pre-injury WM ability in the present. Techniques included age regression to before and after the injury and visualizations of cognitive change and brain plasticity. The active control procedure consisted of non-targeted suggestions from mindfulness meditation practices, involving body and thought awareness with no explicit mentioning of brain injury or working memory-related abilities. Mindfulness meditation has previously demonstrated no or small effects on cognitive abilities in brain injured subjects (Azulay, Smart, Mott, & Cicerone, 2013; Johansson, Bjuhr, & Rönnbäck, 2012; McMillan, Robertson, Brock, & Chorlton, 2002). A number of steps were taken to factor out confounding differences between these two interventions (see Experimental Methods).

An additional 21 subjects were recruited using identical criteria to control for retest effects. They received no intervention and were compensated 1.500 DKK on completion. 19 subjects completed.

	N	TBI / Stroke / other / NA	Age (SD)	Years since injury (SD)	Sex F/M	SHSS: C (SD)
Group A	27	18 / 5 / 4 / 0	45.2 (13.0)	11.6 (11.4)	15/12	7.7 (2.1)
Group B	22	12 / 5 / 5 / 0	47.0 (14.1)	8.0 (6.7)	14/8	6.8 (3.3)
Control	19	4 / 10 / 3 / 2	54.1 (11.7)	7.9 (6.6)	11/8	NA

Table 11. *Descriptives of sample and groups. SHSS:C is the Stanford Hypnotic Susceptibility Scale form C (Weitzenhoffer & Hilgard, 1962). NA = Not available.*

The experiment had three phases (see Figure 18): phase 1 followed by a break (6.7 weeks, SD 1.4) followed by phase 2. Group A received two versions of the targeted procedure. Group B received the non-targeted procedure during phase 1 and the first of the two targeted procedures during phase 2. Each procedure consisted of a weekly session of approximately 1 hour duration for four successive weeks. The control group was passive throughout the experiment but was tested like all other participants.

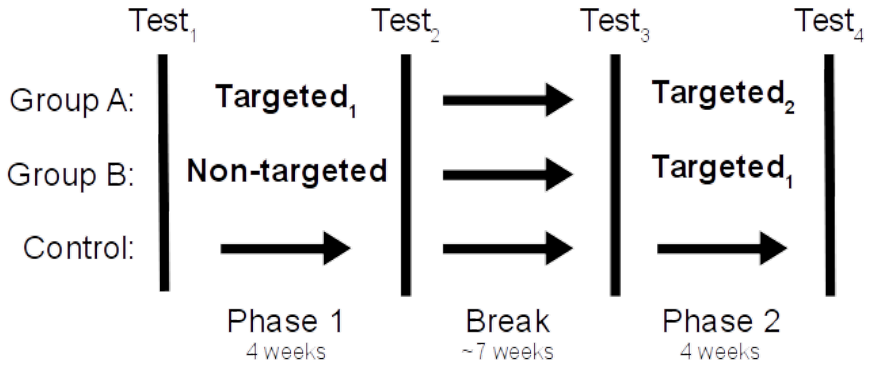

Figure 18. *The experimental design with three groups and four tests (vertical lines). Arrows between tests represents no treatment. The middle column is a break with an average duration of 6.7 weeks. $Test_1$ is baseline and $Test_3$ is follow-up.*

Subjects were tested on the Working Memory Index (WMI) from the Wechsler Adult Intelligence Scale (WAIS-III) and the Trail Making Test part A and B before and after each procedure. WMI is calculated from four tasks: forward digit span, backward digit span, letter-number sequencing and mental arithmetic and has good psychometric properties (Iverson, 2001). All WMI scores were corrected for age and converted to index-100 scores. The B-A Trail Making index is the time cost of alternating between increments in two sequences relative to just incrementing one sequence. It is frequently used as an index of executive control (Sánchez-Cubillo et al., 2009) - the central part of the working memory system. B-A is log-normal so log(B-A) was used as dependent variable.

3.2.2 Results

The results supports the hypothesis that targeted hypnotic suggestion have a distinct positive effect on working memory performance (see Figure 19).

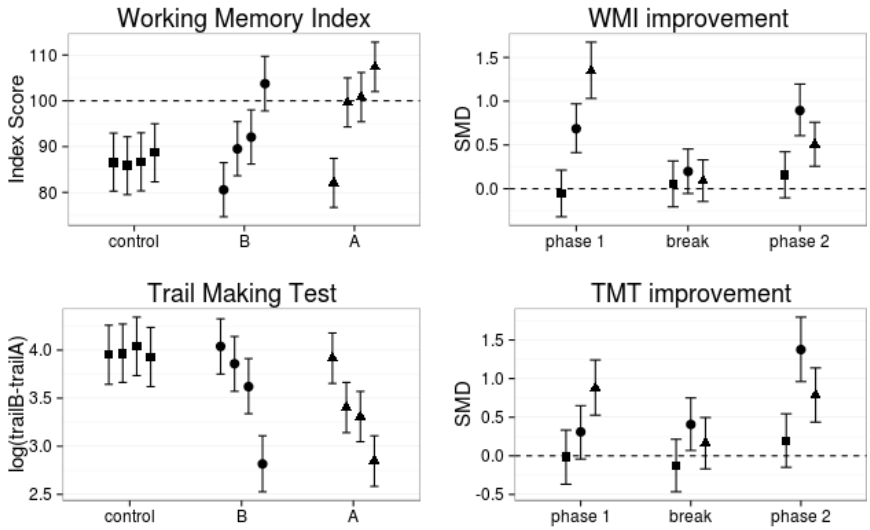

Figure 19. *Per-test and change scores with 95% CI for the Working Memory Index (WMI) and the Trail Making Test. Left: scores for each of the four test sessions. WMI scores in group A and B improved from below (< 85) to above the population average (> 100) following the full program while the control group remained at status quo. In both groups, the major improvements in WMI and TMT occurred during the targeted procedure. The higher the WM capacity and the smaller the TMT time the better. Right: Standardized mean differences for each phase. The control group (squares) is persistently close to zero (dashed line) indicating small retest effects. See results from subtests in Figure 20 and 21 and*

compare to pilot study in Figure 24.

Phase 1: Most importantly, the targeted hypnotic procedure received by group A in phase 1 improved WMI by 0.68 SMD more than the non-targeted procedure received by group B ($p_{LRT}=4.1\times10^{-5}$) with a Bayes factor (BF) of 342 revealing that it is 342 times more likely that group A had superior improvement relative to group B than that they experienced the same effect. Group A improved 1.39 SMD more than the retest effect during phase 1 on the WMI ($p_{LRT}=7\times10^{-13}$, $BF=1.7\times10^{13}$). By convention these estimates correspond to large and very large effects (Cohen, 1992). The selective effect of the targeted procedure was even more apparent for TMT with a similar effect size of 0.63 SMD ($p_{LRT}=0.0007$, BF=37.5) but with no improvement following the non-targeted procedure (SMD=0.30, $p_{LRT}=0.50$) with a Bayes Factor of 6.2 in favor of the null meaning that it is 6.2 times more likely that group B simply experienced the same retest effect as the control group than that group B experienced a separate effect. The retest effect is factored out in all the following unless otherwise specified and a full table of estimates, intervals and inferences can be found in Table 12.

Break: WMI remained unchanged after a 6.7 (SD 1.5) week break following the targeted procedure (SMD=-0.10, $p_{LRT}= 0.46$, BF=7.3 in favor of no difference) as well as the non-targeted procedure (SMD=0.14, $p_{LRT}= 0.76$, BF=9.1 in favor of no difference) when controlling for non-specific and retest effects respectively. This is also true of TMT (see Table 12).

Phase 2: In phase 2, group B had crossed over to the targeted intervention which enhanced WMI by 0.74 SMD ($p_{LRT}=2.0\times10^{-5}$, BF=535) and 1.20 SMD for the TMT ($p_{LRT}=1.2\times10^{7}$, BF=72813). Note that group B improved more on both measures following the targeted intervention in phase 2 compared to the non-targeted intervention in phase 1, thus providing evidence for a distinctive effect of the targeted hypnotic suggestion. Subjects in group A improved another 0.60 SMD ($p_{LRT}=0.0008$, BF=30.0) on the TMT while the evidence was weaker for WMI (SMD=0.35, $p_{LRT}= 0.018$, BF=1.5).

Total improvement: Interestingly, WMI improvement from baseline to the last test in group A (SMD=1.73, 95% CI=1.30-2.17) and B (SMD=1.55, 95% CI=1.13-1.98) had the same magnitude (BF=4.4 in favor of no difference). The same is true for TMT which again demonstrated very large total improvements in group A (SMD=1.83, 95% CI=1.29-2.38) and group B (SMD=2.03, 95% CI=1.47-2.60) with a similar magnitude (BF=2.0 in favor of no difference).

The fact that WMI improved to the population mean indicates that subjects were representative of the normal population before their impairment and that this level was restored following the targeted hypnotic suggestion. TMT improved to better than the

population mean (Nielsen, Knudsen, & Daugbjerg, 1989) with a median improvement from 48 seconds to 24 seconds in the treatment groups.

Robustness: The above inferences were not qualitatively altered when including gender, time since injury, baseline score, SHSS:C score, duration of break, years of education and cause of injury in the model (see supplemental). The robustness to moderating factors indicate that this finding should generalize. For example, participants with an initial performance above the population average improved just as much as those below the normal range.

The reported pattern is evident in each of the WMI subtests (Figure 20). The TMT results were mostly driven by TMT-B, consistent with the hypothesis that the targeted functions were improved whereas the non-targeted visuo-motor skills were not (Figure 21). This experiment numerically replicates a small pilot study on 4 patients (Figure 24). Plots of subtests, additional effect sizes, intervals and inferential results can be seen in Table 12.

3.2.3 Discussion

We conclude that hypnotic suggestion can restore working memory performance in brain injured patients. These data suggest that acquired brain injury does not necessarily impose an irreversible bottleneck on the amount of items that people can effectively store and process, nor the cost of switching between tasks. On the contrary, working memory performance proved to be malleable in a short timeframe.

While these data demonstrate the behavioral effects of targeted hypnotic suggestion, we make no claims about the underlying mechanism. Patients were asked to improve their current working memory functioning based on pre-injury experiences. We do not mean to imply that they literally did this given that hypnosis may in some cases impair recollection (Friedlander & Smith, 1985). We find it more plausible that suggestion facilitated mental imagery or expectations about being unimpaired which may have recruited non-impaired strategies. Other potential mechanisms include unlearning learned helplessness, stimulation of brain plasticity, anxiety-reduction and role-playing. Anxiety reduction is, however, unlikely given that there was no differences in the hypothetized direction on the European Brain Injury Questionnaire as rated by subjects and their relatives (see Figure 22). Narrowing in on possible mechanisms is a topic for future research. Regardless of mechanism, the result is a long-lasting improved working memory performance.

We claim that the improvements were predominantly caused by the targeting of the suggestions which was captured by the difference between group A and B in phase 1 and the break. This view was further supported by the superior improvements in group B during phase 2. The influence of confounding factors was ruled out by design and data. The design

controlled for retest effects, relaxation responses and expectancy effects. Demographic parameters did not explain the effects. Furthermore, the results are not readily explicable in terms of specific low-level strategies since the intervention did not mention or resemble the outcome measures and the results generalized to dissimilar visual and verbal material. The two treatment groups had the same level of motivation and belief in the effectiveness of the hypnotic procedures throughout the experiment (see Figure 23).

Suggestibility did not influence the effects notably, contrary to what is commonly found in the hypnosis literature (Halligan & Oakley, 2013; Oakley & Halligan, 2013; Kirsch & Braffman, 2001; Raz, Shapiro, Fan, & Posner, 2002). However, suggestibility has been found not to be a precondition for effectiveness in surgical patients (Montgomery, David, Winkel, Silverstein, & Bovbjerg, 2002). Suggestion for health improvements may be easier than captured by suggestibility scales, or may work through different mechanisms than suggestibility as measured by classical tests. Again, we make no claims about the underlying mechanism of change. Future studies using non-hypnotic suggestion on patients and hypnotic suggestion on healthy subjects would be ideal to determine the preconditions and scope of the effect.

The novel contribution of the present work is the demonstration that targeted hypnotic suggestions causes a very large and long-lasting improvement in human working memory-related performance. In particular, the results give evidence for the hypothesis that hypnotic suggestion can be used to influence working memory, and furthermore, that this is more efficient than known bottom-up behavioral and biological approaches.

3.2.4 Experimental procedures

Controlling nuisance factors: The following steps were taken to ensure that the contrast between the targeted and non-targeted procedures was pure: (1) Procedures were written up as a manual as dictated by the hypnotist. (2) Manuscripts for the two procedures were identical with respect to the induction and termination of the hypnotic state, and accordingly, only differed in the targeting of the suggestions. (3) Duration of the sessions and all information material were kept equal. (4) testers were blinded and participants were blinded to the existence of any other groups in the study. (5) To homogenize expectations, all participants were informed about a successful pilot study (see supplemental) but that the outcome of the current experiment was unknown. (6) To minimize test-specific effects neither procedure contained any training or mention of the neuropsychological testing situation and materials.

Inference model: Results were modelled using a mixed effects model with a covariate for retest, non-targeted, and targeted effects for each phase and a random intercept per subject. This model explicitly controls for retest and non-targeted effects when assessing targeted

effects. Covariates were tested against the null using the likelihood ratio tests (LRT) where the null model was the full model less the covariate(s) in question. The strength of evidence for the covariate being zero was quantified using the Bayes Factor (BF) with a Cauchy(0, $\sqrt{(2)}/2$) prior on covariates (Rouder & Morey, 2012). The Bayes Factor is the shift in the posterior odds ratio between the full model and the null model that was brought about by the data. Conventionally, a BF of 3-20 is labelled as "positive evidence", 20-150 is "strong evidence" and >150 is "very strong evidence" (Kass & Raftery, 1995) but we invite readers to simply interpret each BF as an odds ratio rather than transforming them into crude universal categories. As a general rule, these labels are more conservative than p-values (Wetzels et al., 2011). An advantage of Bayes Factors relative to p-values is that they quantify evidence for both the full and the null model while accounting for model complexity (Posada & Buckley, 2004). Sensitivity analyses using different Bayes factors are available in the supplementary material.

For parameter estimates a more informative Normal(0, 0.5) prior was used which expresses a skeptical prior belief that there is around 98% probability of the effect size being less than 1 SMD. Therefore the effect sizes reported here are smaller than had they been obtained by conventional maximum-likelihood estimation. With an uninformative prior, the means and CIs coincide largely with those obtained from maximum-likelihood methods.

3.2.5 Acknowledgements

This work was supported by the European Research Council and the Karen-Elise Jensen foundation. All authors contributed equally.

3.2.6 Supplemental data

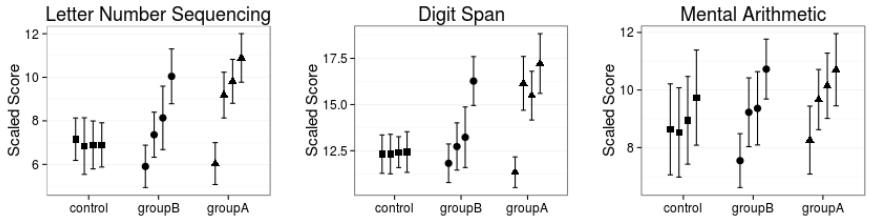

Figure 20: *Means and 95% bootstrapped confidence intervals for each WAIS-III working memory subtest, corrected for age. Digit Span is a compound score for forwards and backwards digit span. The overall pattern from figure 19 is well representative of the individual subtests.*

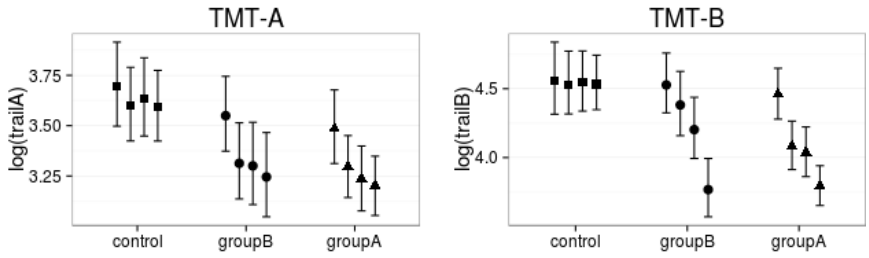

Figure 21: *Means and 95% bootstrapped confidence intervals for each TMT subtest. The B-A index is mostly driven by TMT-B as subjects improved approximately equally in TMT-A.*

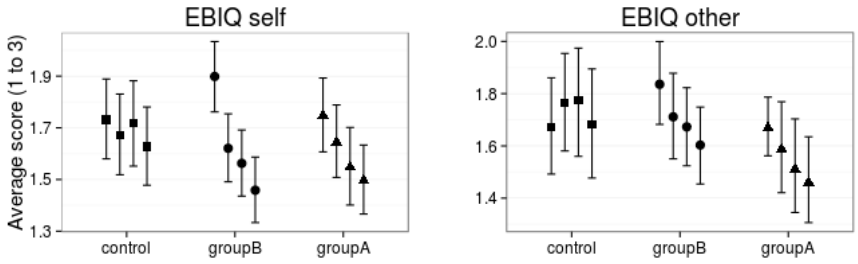

Figure 22: *Average scores and 95% bootstrapped confidence intervals on the European Brain Injury Questionnaire (EBIQ) as rated by the patient self (left) and a significant other (right). Most items on EBIQ could be expected to covary with anxiety so the average score can be used as a crude assessment of the hypothesis that anxiety reduction explains the improvements on the outcome measures. There is no indication that group A experience fewer problems on the EBIQ than group B. To the contrary, patients in group B reported superior improvements during phase 1 on the EBIQ, thus discrediting the anxiety-hypothesis.*

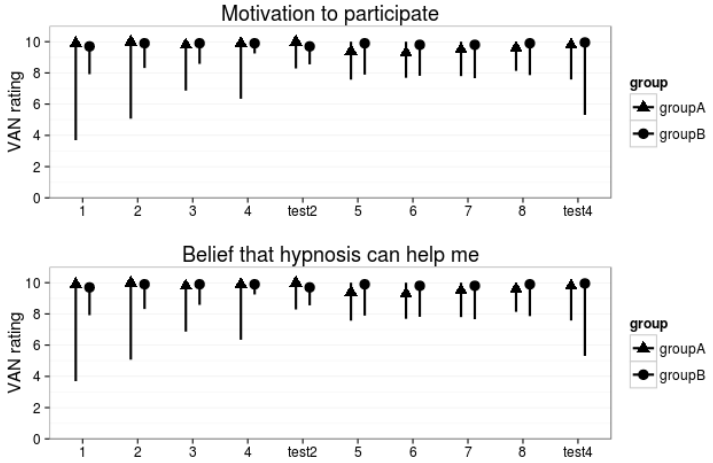

Figure 23: *Medians and interquartile intervals for the Visual Analogue Scale rating of motivation to participate and belief that hypnosis can help. There is a clear ceiling effect and no apparent differences between groups. To minimize bias, these ratings were filled out privately by the participants before each treatment and put in an envelope which was opened when blinding was lifted, i.e. after all participants had completed.*

3.2.7 Supplemental: Pilot study

This paper replicates findings from a pilot study using the same design. The pilot results are shown in figure 24. Briefly, we recruited 4 patients and divided them into a group A and group B so that base WMI score was equal in the two groups. The group B intervention in phase was listening to relaxing music between induction and termination and no additional suggestions were given. The experimental design was identical to the current study but participants in group B were not given a test after the break.

Group A improved by 20 points following the targeted suggestions in phase 1, stayed relatively stable during the break and increased further in phase 2. Group B stayed at status quo during the control procedure in phase 1 and had improved 20 points following the targeted suggestions in phase 2. The present study not only replicate these finding conceptually, but also quantitatively.

The "pure state" procedure was eventually altered into the much more closely matched non-targeted mindfulness-like procedure that was administered to group B during phase 1 in the present study.

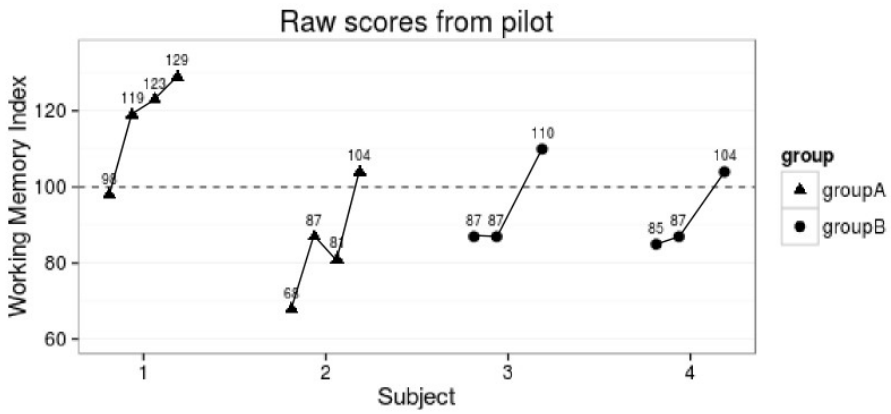

Figure 24: *Data from the four participants in the pilot study. Group A (blue) and group B (red) essentially followed the same pattern and the same magnitudes as observed in the present study. The dashed line is the average of the normal population. Compare to figure 19.*

	dep	null	id	group	phase	cov	MD	SMD.ml	p_LRT	SMD	SMD.low	SMD.up	SMD.p.low	SMD.p.up	SMD.p.pos	BF_BIC	BF_g	BF_PS	ESS_PS
1	WMI	-C1	1	control	1	1	-0.856	-0.069	0.6405901	-0.101	-0.396	0.196	-1.077	0.914	0.419	14.8*	7.5*	5.4*	3933
2	TMT	-C1	1	control	1	1	-0.002	-0.003	0.9873677	0.014	-0.404	0.448	-1.457	1.445	0.509	16.5*	5.2*	4.8*	5753
3	WMI	-Cb	2	control	break	1	1.632	0.132	0.3874879	0.103	-0.191	0.400	-0.893	1.093	0.582	11.4*	6.6*	5.3*	4660
4	TMT	-Cb	2	control	break	1	0.040	0.063	0.7398428	0.108	-0.316	0.535	-1.329	1.583	0.559	15.6*	6.6*	4.1*	6932
5	WMI	-C2	3	control	2	1	1.053	0.085	0.5769568	0.114	-0.182	0.410	-0.885	1.105	0.588	14.1*	8*	5.1*	7054
6	TMT	-C2	3	control	2	1	0.025	0.040	0.8331666	-0.058	-0.475	0.360	-1.544	1.375	0.468	16.1*	6.7*	4.5*	11428
7	WMI	-BC1	4	B	1	1	9.358	0.756	0.0002067045	0.813	0.400	1.233	-0.221	1.853	0.941	59.2	55.5	332.9	45
8	TMT	-BC1	4	B	1	1	-0.101	-0.159	0.5029062	-0.281	-0.859	0.282	-1.779	1.234	0.355	13.2*	3.5*	2.2*	3392
9	WMI	-BCb	5	B	break	1	0.778	0.063	0.7627227	0.095	-0.305	0.498	-0.947	1.117	0.571	15.8*	8.7*	4.7*	5558
10	TMT	-BCb	5	B	break	1	-0.306	-0.481	0.06393477	-0.576	-1.141	-0.004	-2.089	0.926	0.222	3*	1.3*	2.3	1692
11	WMI	-BC2	6	B	2	1	11.220	0.907	0.0000204047	0.855	0.434	1.281	-0.190	1.895	0.949	529.9	534.2	1077.1	376
12	TMT	-BC2	6	B	2	1	-0.903	-1.418	0.0000001158185	-1.509	-2.118	-0.910	-3.032	-0.001	0.023	76264.8	72063.4	169152.9	118
13	WMI	-ABC1	7	A	1	1	9.433	0.762	0.00004074493	0.726	0.339	1.121	-0.311	1.751	0.920	274.6	335.1	244.9	86
14	TMT	-ABC1	7	A	1	1	-0.473	-0.742	0.000739525	-0.738	-1.279	-0.193	-2.256	0.734	0.164	18	35.5	13.1	473
15	WMI	-ABCb	8	A	break	1	-1.742	-0.141	0.4609052	-0.132	-0.502	0.244	-1.149	0.894	0.398	12.6*	7.3*	4.1*	8021
16	TMT	-ABCb	8	A	break	1	0.233	0.366	0.1234735	0.364	-0.161	0.897	-1.128	1.848	0.687	5*	2.2*	1.4*	5805
17	WMI	-AC2	9	A	2	1	5.873	0.475	0.0178328	0.432	0.041	0.824	-0.594	1.462	0.798	1	1.5	2.4	3236
18	TMT	-AC2	9	A	2	1	-0.538	-0.844	0.0007496957	-0.846	-1.409	-0.294	-2.340	0.660	0.131	17.8	21.9	30.7	1526

Table 12:. *Effects estimates and inferences. A = group A, B=group B, C=control. "-ABC1" is read as "The effect of A in phase 1 when the effect of B and C in phase 1 is subtracted". SMD = Standardized mean difference. SMD.p = posterior predictive of the SMD. SMD.p.pos=probability that the posterior predictive is positive, i.e. larger than zero. This is a useful index for how likely you are you are to observe these effects in real life. For all targeted effects but TMT-AC2 the posterior predictive* BF_{BIC}*=Bayes Factor with unit information prior on covariates, derived from the BIC of the ML model.* BF_g*=Bayes Factor with g-prior.* BF_{PS}*=bayes factor with Normal(0, 0.5) prior inferred using the product space method. * = BF in support of the null.* ESS_{PS} *is the effective sample size for* BF_{PS} *and it should be above 500 before* BF_{PS} *begins to get some reliability.*

3.2.8 Supplemental Experimental Procedures

Intervention procedures

Induction of hypnosis: The same induction was used in both procedures. Participants were asked to increasingly relax and to pay attention only to the hypnotist's voice. Participants were asked to prepare to work towards improvement during the session. Inductions typically lasted 5-10 minutes and made use of visual imagery, e.g. walking down a staircase or going down with an elevator.

Targeted procedure: The hypnotist told the participants that they could improve their memory and attentive functions. The participants were asked to perform various tasks during the session:

- *Imagining previous strategies:* Participants were asked to experience how they solved problems and remembered things in their lives prior to the injury.
- *Imagining the process of recovery:* Participants were asked to imagine the physical growth of neurons and connections in their brains.
- *Imagining results of recovery:* Participants were asked to feel what it would be like to be able to use their previous abilities and strategies in their current life situation.
- *Trusting improvement:* Participants were asked to believe that it is possible to make improvements regardless of what they had been told after hospitalization or rehabilitation.
- *Goals:* Participants were asked to identify their current problems and what they would like for the future in terms of their own cognitive abilities.

Translated excerpts from session 1: *"While you are fully relaxed, I will ask you to create a picture of your life history as a line (...) once it was easy for you to concentrate… without getting tired as you do now… So if you are not already there, then go... back in time, longer and longer into your own past… to the time before the incident… just relax, it does not have to be experienced in one particular way… just accept the way you experience this (...) Now find three situations… in your past, before the incident… where you concentrate and remember all that is relevant to you with ease and without any trouble… something that you today find hard to do (...) Now, take your re-found knowledge of how it used to be… and move forward… towards now (...) if something is blocked or closed… you just find another way… and a new way can be as good… sometimes even better… than an old one (...) Now move to the future… and experience yourself in a situation… where you can do something you thought was impossible… experience it very life-like… notice as many*

details as you can."

Non-targeted procedure: The hypnotist used suggestions known from mindfulness meditation. In particular, the suggestions including focused attention on bodily sensations, e.g. breathing, and open monitoring of the stream of thought.

Translated excerpts from session 1: *"Feel that you sit in the chair… notice which part of your body occupy your attention… (…) perhaps you will feel the sensations change (…) also feel your tendency… to want to be with some impressions more than others (…) ask yourself what you experience right now… do not judge what you feel… just feel it (…) extend your attention to include the fact that you are just sitting here in silence… the silence and closeness… right now and here… just sensing what may come… without judging (…) Feel what it is like to breathe… notice that you chest rises and lowers… feel in the inhale and the exhale (…) Now direct your attention to your left foot… from the toes to the heel… if you have tensions… let them go (…) maybe you feel sensations or emotions… when you feel something… be attentive, curious and open to what you feel.*

Termination of hypnosis: The same termination was used in both procedures. Terminations were brief as the hypnotist slowly counted backwards from 10 to 1 while suggesting the participant to wake up in a pleasant way, and to prepare to continue their day.

Recruitment of subjects

Patients were recruited from Danish brain injury communities based on self-reported working memory problems, such as poor concentration and forgetfulness. The cognitive nature of the self-reported problem was subsequently confirmed in patient records as well as neuropsychological tests (including MMSE and clock drawing) and relevant aetiology. Exclusion criteria were psychiatric or neurodegenerative disease and/or other severe cognitive dysfunctions (e.g. severe aphasia) and concurrent rehabilitation at any time during the study. Heterogeneity in the sample was intentional since the hypothesis is invariant with respect to aetiology and demographic variables. See table 11 for a description of the sample.

Subjects were continuously enrolled and randomly allocated to two groups using coin tosses.

Model and inferences

Regression model: Observed data were modeled using a mixed effects model with retest effects (observed in the control group), non-targeted effects (additional effect observed in group B) and targeted effects (additional effect observed in group A) as fixed effects. These effects accumulated over time so that they reflect change scores from the previous test session to the current. Expressed mathematically:

$$score_{g,s,i} = intercept + X_s \sum_{j=1}^{s-1} (retest_j + X_{g,n}\, nontargeted_j + X_{g,t}\, targeted_j) + subject_i$$

… where g is the group index, s is the test session index and i is the subject index. X is design indicators where $X_s = 0$ when $s = 1$ (baseline) and $X_s = 1$ otherwise, $\{X_{g,n} = 0, X_{g,t} = 0\}$ for the control group, $\{X_{g,n} = 1, X_{g,t} = 0\}$ for group B and $\{X_{g,n} = 1, X_{g,t} = 1\}$ for group A. However, when $s = 4$, i.e. phase 2 or test session 4, $\{X_{g,n} = 0, X_{g,t} = 1\}$ for group A because no non-targeted effects were assessed directly. This is due to the non-crossover design, which was intentional since strong order effects would be likely if group A should transition from the targeted to the non-targeted intervention.

Subject was added as stochastic part to account for dependencies between repeated measures. $subject_i$ is the subject-specific offset relative to the population baseline captured by the intercept. The full model had a good fit (r^2=0.896, random effects included)

Inferences: the specific parameter estimates are of primary interest. Thus we did not use more general procedures such as testing the group x session interaction in a repeated measures ANOVA because they answer much more general questions ("do the two groups of individuals develop differently?") than would be informative here. Likewise, latent change scores (McArdle, 2009) and other growth models might better represent the actual time course of the sessions but at the cost of a much higher model complexity than is needed to answer the main research questions of this paper.

Bayesian and frequentist inferences: Likelihood Ratios (LR) and Bayes Factors (BF) both yield posterior odds of the two models in question. They do differ in that the LRT considers only the maximum likelihood estimates (MLE) against the null whereas the BF considers the full set of credible parameter values. By so doing, BF inherently penalizes model complexity whereas LRT has to make an additional decision based on Akaike

Information Criterion (AIC) or Bayesian Information Criterion (BIC) to avoid overfitting.

For this particular model, the frequentist and Bayesian analysis yield almost identical parameter estimates and intervals. This is also true for the BIC-based Bayes Factors (Wagenmakers, 2007), $BF_{BIC} = \exp(\frac{BIC_{full} - BIC_{null}}{2})$, g-priors-based Bayes Factors (prior is Cauchy(0, sqrt(2)/2)) on SMD and on Product Space (PS) derived Bayes factors where we used a Normal(0, 0.5) prior on covariates. These three Bayes factors were used to assess the sensitivity of the inferences to prior specifications.

The BIC may be interpreted as the unit information prior which is as informative as one data point. The unit information prior is very wide (uninformative) and thus puts less mass on the null than the g-priors and the related JZS prior which has 95% mass between -9 and 9 SMD. Thus Bayes factors based on BIC is more supportive of the null than those based on the g-prior because the relative change from prior to posterior probability at the null is larger. Likewise, for the PS Bayes factor we set a prior 95% probability between -1 and 1 SMD and is thereby the most informative prior with the least supportive of the null and conversely similarly larger supportive of the alternative. Analysis using these three priors are listed in table 12 and can be explored using the accompanying data analysis script.

As mentioned in the main text, we find inferences about model odds given the current dataset much more informative than p-values which are simply a decision criterion to control the type-1 error rate in the limit of infinite perfect replications and does not quantify the probability of type-1 error rate in a single experiment.

Effects: Effects were derived as means of the posterior and 95% highest posterior density regions, also called credible intervals. A Normal(0, 0.5) prior was used on covariates which puts 95% probability mass between -1 and 1 SD. This causes the effect sizes to be around 10% smaller than maximum-likelihood estimates reflecting a somewhat skeptical prior view on big effect sizes. See table 12 for effects (retest is relative to intercept, non-targeted is relative to retest, targeted is relative to non-targeted+retest), effect sizes (SMD = difference / $SD_{pretest}$, credible intervals, p_{LRT} and Bayes factors. Notice that all retest effects and break effects lend support to the null hypothesis that the effect is zero while the treatment effects led support to the alternative.

Nuisance covariates: were tested using a likelihood ratio test between (1) the mixed effects model described in Methods and Materials plus the interaction between the covariate and the fixed effect in question and (2) this model without the fixed effect. Effectively, this models the extent to which the covariate can explain the variance otherwise caught by the fixed effect, i.e. the extent to which the covariate is an alternative interpretation or a confounder to the effect of the behavioral intervention. No covariates qualitatively replaced the fixed effect (see main text). All covariates were mean-centered.

Notes on specific covariates:

- *Cause* was tested on Traumatic Brain Injury and Stroke patients only (N=54), ignoring other patients. There were too few cases of other causes (e.g. cerebral hypoxia or encephalitis) to model them.
- *SHSS:C* (suggestibility) was only tested on group A and B (N=48) and not the control group, since the control group was completely hypnosis-free.
- *Break duration* was only tested on the interactions with the three fixed effects for the break, i.e. change from session 2 to 3.
- *Education* was the nominated number of years between public school and the latest completed education. A value of 0 corresponds to 9 years in public school.

3.3 Epilogue: mechanisms and implications for theory

We have kept the paper almost entirely free from theoretical considerations in order to present the findings as they are. Simply establishing that the targetedness of hypnotic suggestion causes something that seems to be a genuine high-level cognitive improvement is a major step in and of itself. It is therefore also one that should be subjected to scrutiny before making too much of it theoretically. Such scrutiny includes replications and extensions. I therefore consider the following discussion to be a catalogue of ideas for future studies rather than definite conclusions.

3.3.1 What was improved?

We observed improved performance on outcome measures typically thought to represent working memory functioning. It may seem obvious to conclude that working memory per se was improved. However, there are several other explanations which offer the same end result but through different underlying changes. I will consider the merits of each alternative in the following.

First, we can derive that the improvement is a cognitive improvement. Given that the patient was the only common factor between the hypnosis sessions and the testing sessions and that differences in the hypnosis sessions affected the results in the testing sessions, the mediating factor must be a property of the patient. Given that the targetedness is a semantic property of hypnotic suggestion (the targetedness) and that decoding semantics is an entirely cognitive phenomenon, the mediating factor must be cognitive. From here, Morgan's canon dictatet that low-level alternative explanations should be considered. I list such alternatives here in the approximate order of explanatory level.

Chance: It could be that the findings of the hypnosis study is simply due to chance - an extreme 1 in 10,000 study which would fail to replicate the other 9,999 times. Such a claim could be substantiated by noting that the findings are at odds with many previous findings. First, it has never been published before. If the effect is as versatile as our study seems to suggest, someone should have noticed and published it after decades of basic and clinical hypnosis research. Second, persistent positive effects of this magnitude following such a short intervention have not hitherto been reported in patients with acquired brain injury in the chronic phase. Third, as far as I know no established theory of hypnosis allow for hypnotic suggestion to work on high-level cognition (see discussion of hypnosis theories in a later section).

Independent replication would be the primary test of the chance-explanation. We have replicated the results under various conditions. (1) in the pilot study, Here, patients had EEG mounted simultaneously. (2) In group A during phase 1. The script had been rewritten relative to the pilot study. (3) in group B during phase 2, in which case subjects were arguably more familiar with the hypnotic procedure and had already been through a mindfulness-hypnosis. (4) We have casually run two more subjects through a 4-session treatment with a once-again rewritten script, and observe effects of the same magnitude. Finally, the persistency of the effect following the break in both the pilot and for group A count as replications with respect to the chance-claim that the results were merely lucky momentary fluctuations in the state of the patients. Given that the results were replicated under these differing conditions, the claim that there was a consistent methodological bias is at least somewhat ameliorated. The latter would be most forcefully tested by an independent replication.

We set a prior bayes factor of 1 between the full and null models, meaning that we assumed equal prior probability of the presence and absence of a unique effect of targeted hypnotic suggestion. One could formalize the "chance" skepticism by favoring the null model a priori say by setting a BF_{null}=10, corresponding to 9% prior probability of the presence of an effect. The critical bayes factor (difference between group B and group A in phase 1) was 342 for the working memory index and 37.5 for the trail making index. With the skeptical prior, the posterior bayes factors would be 34.2 and 3.75 - still in favor of the effectiveness of targeted suggestion. One would need to set the prior bayes factor at around 100 in favor of the null to stay unconvinced by all results, even after having observed the results from the pilot study. Given that the pilot study constitute the only direct prior evidence, I find the latter skeptical prior hard to defend.

Test-specific learning: It could be that the targeted suggestion caused improvements specifically on the test material but nothing else, i.e. a non-generalizing effect. This explanation would predict smaller- or no improvements on tests which we did not administer. Therefore it cannot be supported or falsified from the present dataset.

None of the suggestions in the targeted condition mentioned the specific tests, test performance, or the test context. For the test-specific explanation to work, subjects would have to be very goal-directed in using the suggestions to improve on the tests only, e.g. by staying focused in the test situation or a particular test throughout each hypnosis session and interpreting all suggestions as pertaining directly to that. But since patients initially set specific goals for the everyday problems they wanted to improve and since the testing was for research purposes only, I find this unlikely.

Context-specific effect: Hypnosis and testing took place in the same building. For around half of the patients, hypnosis and tests took place in the same room. It could be that the effects of the hypnosis sessions were tied specifically to that building/room. We can discard the room-explanation since every single patient in group B improved, so if around half of them were tested in a different room, it must have generalized across rooms. Unfortunately, we did not register which room patients were tested in to quantify such an effect. With respect to the building, we know that at least some of the effect "leaked" out of the building since patients and relatives observed a positive effect in everyday life.

Even if the full effect could be attributed to the room- or building-hypothesis, the results still demonstrates that the targetedness of hypnotic suggestion improved performance on the working memory-related tests.

Short-term memory: One could also posit that the effect can be attributed to something lower in the cognitive hierarchy than working memory, e.g. simply the slave systems of Baddeley's working memory model. Such a claim would be at odds with the way processing is isolated in the Trail Making B - A index and the letter-number sequencing task. Extensive manipulations of the sensory input have to take place to produce a correct answer for these tasks.

Changed criterion: Hypnotically enhanced recall has been shown to do little more than to decrease the response criterion of the subject while the underlying ability is unchanged (Dywan & Bowers, 1983; Klatzky, 1985). I.e. the subject recalls more details correctly but also makes more mistakes. This cannot be the case in this experiment since the subjects would not benefit from a changed criterion alone in neither of these outcome measures. Trail Making is stopped and rewinded on mistakes (a time consuming event, thus yielding poorer scores) and the full trial is scored as failed upon a single mistake in in the Working Memory Index.

Reduced anxiety: Performance could have been improved simply because subjects are less anxious after having heard reassuring statements in the targeted hypnosis sessions. In the test situation, less anxiety could be said to increase the focus on the task at hand by reducing the attention allocated to monitoring one's own emotional state or intruding thoughts about being inferior. According to this explanation, changes in anxiety and test performance should covary.

We have data that indirectly inform us about the merits of the anxiety-explanation. We administered the danish version of the European Brain Injury Questionnaire (EBIQ, Teasdale et al., 1997) to patients and significant others at all four tests. EBIQ contains 62 items about somatization, cognition, motivation, emotion etc. and asks how much each was experienced during the last few weeks. Possible answers are (1) not at all, (2) a little, and (3) a lot. Anxiety should affect most of these items, perhaps except the somatization items.

The results runs somewhat counter to the predictions of the anxiety-theory (see Figure 25). There is no sign of group A showing superior anxiety reduction during phase A or group B showing superior anxiety reduction in phase B (where the change is comparable to the control group), either as rated by the patient himself/herself or by a significant other. In fact, there was the opposite trend in the self-score during phase 1, indicating a dissociation between subjective experience and objective performance.

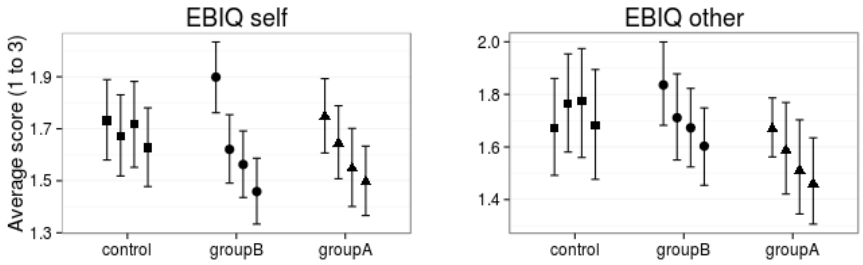

Figure 25: *Score on the "core" index of the European Brain Injury Questionnaire as rated by the patient (left) and a significant other (right). Although there are improvements in the two intervention groups overall, they do not follow the pattern from the objective scores.*

While our main experiment was ongoing, I became aware of two anxiety-focused papers on hypnosis and brain injury which resembled our experiment. Sullivan (1974) stratified 24 young subjects with acquired brain injury into three groups. The intervention consisted of 4 weekly hypnosis sessions of which the first three was simply hypnosis tests on a modified version of the Stanford Hypnotic Susceptibility Scale to familiarize the subjects with hypnosis. The fourth and critical session consisted of a short passage with suggestion on feeling relaxed and safe. The aim was to reduce "catastrophic anxiety". An active control group received 4 weekly sessions with a (non-hypnotic) systematic relaxation intervention. There was a passive control group as well.

Improvements of SMD=0.45, 0.27 and 0.05 were observed in the treatment, active control, and control groups respectively on a picture completion task[35]. Another outcome measure showed no general improvement in the treatment group but a differential improvement between high- and low-susceptible patients. Sullivan concluded that hypnosis could reduce anxiety sequelae of brain injury, which in term improved behavior on tests.

35 Sullivan only published variances for change scores. Variances from Gray et al. (1992) were used for standardization. The brain injured sample from Gray et al. had a similar age and was tested in the same time period.

An older experiment by Fromm et al. (1964) made an almost inverse experiment. Fromm hypnotized 9 healthy subjects to think that they had sustained a brain injury. Blinded testers rated subjects as more brain injured under this condition than three other control conditions, although the ratings was not as strong as would be expected for patients with actual brain damage. Fromm also concluded that the hypnosis induced "catastrophic anxiety" and that this constitutes a part of the behavioral deficits observed in brain injury patients.

Neither Fromm (1964) nor Sullivan (1974) actually assessed the anxiety level of their subjects and therefore their conclusions about catastrophic anxiety as mediating factor does not follow directly from the data. What they do demonstrate is that *suggestion about catastrophic anxiety* increased or decreased performance on neuropsychological outcome measures. This, in turn, is of course indicative that anxiety might be a mediating factor in this experiment. If Sullivan and Fromm are right in this assumption, then we have established that a reduction in anxiety can increase cognitive performance in situations like our experiment. But we saw an indication of a superior anxiety reduction in the mindfulness-suggestion condition compared to the targeted suggestion condition. It is possible that different mechanisms are at play: that mindfulness-suggestion primarily increased performance via anxiety reduction whereas targeted suggestion primarily increased performance through one of the other mechanisms mentioned in this section.

Although our results seem to indicate that anxiety-reduction cannot account for the observed effect of targeted suggestion, it is certainly not ruled out. If anxiety indeed was the underlying cause, this explanation would still predict that the effect should generalize.

As a side note, Wagstaff et al. (2001) did an experiment similar to that by Fromm et al. Wagstaff et al. tested 60 students in a study on amnesia test malingering. Here, hypnotically induced "brain injury" yielded behaviors which resembled malingering more than organic brain injury. Wagstaff concluded that hypnotically induced "brain injury" was not a realistic model of real organic brain injury. Similarly, Fromm et al.'s "brain injury models" did not show as gross deficits as would be expected from organic brain injury. It seems that hypnotically induced brain injury models are not realistic and therefore cannot be used to study the psychological mechanisms of suffering a brain injury.

Role-playing or expectancy: So-called "non-state" theories of hypnosis posit that the mechanisms that bring about hypnotic behavior are mundane and that hypnotic behavior in this sense is no different than responses to other kinds of suggestion. These theories include that the patients were role-playing to be healthy, role-played to be good experimental subjects, and that they did well on the tests because they expected to do well. These theories are meant as a challenge to the view that hypnosis involves special processes. That discussion does not have much relevance for the interpretation of our results. In either case, the patient actually succeeded in improving their performance which had hitherto been impaired.

Improved working memory: It could be that working memory per se improved. Working memory is often thought to be a multi-component system (c.f. Baddeley, 2007; and Cowan, 1988), so this explanation is mute with respect to the exact distribution of improvement across components as long as no single component is hypothesized to be worse than what would be compatible with the observed data. In the present case, there was improvement both on auditory span + processing (WAIS) and visual task switching, so it must at least involve the central executive since there is processing in both of these tasks which cannot be carried out by the slave systems / short term memory.

That working memory improved per se is not contradicted by design or data. It is however at odds with prior knowledge. I am not aware of any 4-to-8 hour intervention that resulted in long-term working memory improvement. This is why I have made an effort to consider all alternatives in this epilogue and why I embedded some prior skepticism in the prior on the covariates of the statistical model. I used a Normal(mean=0, sd=0.5) prior which puts around 95% probability on an SMD between -1 and 1. It caused the SMD to shrink around SMD=0.2 towards zero for both group A and B relative to a maximum-likelihood estimate (see table 12 for a comparison of ML-effect sizes and bayesian effect sizes).

Improvement of other high-level functions: It could be that other high-level cognitive functions improved but that working memory did not. There is a great overlap between the so-called "higher cognitive functions" among which working memory belongs. For example, it has been speculated that fluid intelligence is entirely a property of the central executive component in working memory (Engle et al., 1999. See discussion in Conway et al., 2002). The central executive is also said to govern attention (Baddeley, 2007; Cowan, 2005) and to play a central role in resolution of cognitive conflicts (Botvinick, 2001). Executive functions is a rather vague term referring to e.g. planning, problem solving, task switching and other factors in mental flexibility. Such abilities covary with working memory functions so it is tempting to say, that working memory plays a role here as well (see also Conway, 2002; Conway, 2003; Alloway, 2010; Kane, 2004, for similar arguments that working memory is intertwined with most high-level cognition).

Careful experiments can isolate these constructs through multiple contrasts. But any given test will share variance from many of these constructs, thus not allowing certainty about which constructs contributed how much to the behavioral improvements. Here we have to use reverse inference: (1) if we have prior knowledge that a given outcome measure shares more variance with working memory than other constructs and (2) we observed an improvement on the outcome measure, then it follows that (3) working memory has the highest posterior probability of having improved although other underlying causes are still probable.

As a reminder, I use this “cognitive language” as a reference to relationships between how subjects perform on a range of tasks which share more or less variance with other tasks. For high-level cognition, there is a great overlap between the categories, i.e. the transfer domains, making it different to make an empirical distinction between them.

3.3.2 Suggestibility and clinical effects

It is a common finding in the hypnosis literature that the degree of response to suggestion covaries with suggestibility as measured by various scales. We used the “golden standard” Stanford Hypnotic Susceptibility Scale form C (Weitzenhoffer & Hilgard, 1962), but did not observe any relationship between suggestibility and improvement (see figure 26). As mentioned in the hypnosis paper, this leaves us with two possibilities. Either the treatment suggestions are so easy that SHSS:C does not discriminate between them.

Alternatively, the effect was not driven by the same mechanism as that of the basic sciences literature. In contrast to the cognitive neuroscience literature, suggestibility is rarely assessed in the clinical literature. For example, without assessing suggestibility, successful results were obtained by using hypnosis to reduce discomfort in patients with irritable bowel syndrome (Moser et al., 2013), reduce side effects following breast cancer surgery (Montgomery et al., 2007), or reduce stress during conscious sedation (Faymonville et al., 1997). The relatively large effect sizes in these studies imply that effects are not driven by a subgroup of patients. While most clinical studies who assess suggestibility do find a ranking of treatment effect with low-to-high suggestible patients, some do not (see e.g. Montgomery et al., 2002; Askay et al., 2007). It seems that being highly suggestible a positive predictor for the effectiveness of the intervention but not a necessity.

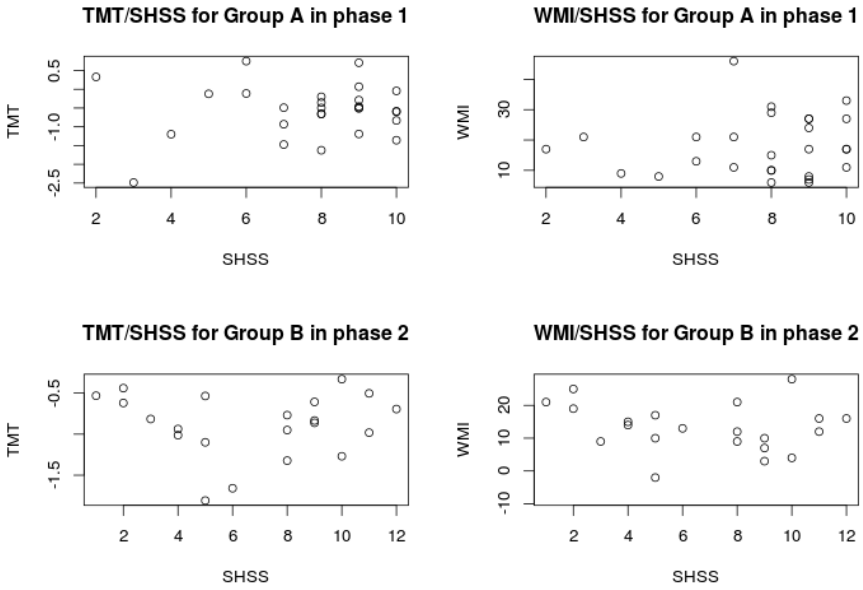

Figure 26. *Changes in outcome measures (ΔTMT and ΔWMI) as a function of baseline SHSS score for phases with targeted suggestions. There is no relation, i.e. patients' improvements are not limited or facilitated by suggestibility.*

3.3.3 Implications for hypnosis theory

Our experiment was not designed based on any theories of hypnosis, so it did not directly assess necessary variables to directly distinguish between predictions of different theories. Still, it remains vaguely informative about at least some theories.

Suggestibility and executive ability: The finding that suggestibility was the same or higher in patients than in healthy subjects does lend support to the view, that hypnotic behavior is not bottlenecked by higher cognitive abilities such as mental flexibility, intelligence, central executive etc. These findings are contrary to the predictions of Neodissociation Theory (Hilgard, 1974, 1977), which says that executive resources are used to maintain the dissociations - at least to the extent that the executive is loaded enough for its capacity limit to play a role in the hypnotic behavior resulting from this upheld dissociation. The findings are more in line with Dissociated Control theory (Miller & Bowers, 1993; Woody & Bowers, 1994) and Cold Control Theory, which posits that hypnotic behavior and experiences are caused by a reduction in executive control[36] which

36 The supervisory attentional system in Norman & Shallice's information processing model, to be exact. It is not entirely clear from Miller & Bower's theory how conflicts between competing

allows suggestion to trigger behavior more directly without the usual corrections by a reality testing executive.

Hypnotic meditation? Our group B did meditative practices during the hypnotic procedure. To my knowledge, this concurrent hypnosis and meditation has never been done before, although several studies have looked at relationships between hypnotic performance and meditative performance in the same individuals (see e.g. Nuys, 1973). Such a condition is interesting as seen from the cold control theory and socio-cognitive account of the difference between hypnosis and meditation, which under these theories can be said to be opposite phenomena (Semmens-Wheeler et al., 2012; Dienes et al., in press). In these theories, hypnosis is associated with an inaccurate meta-awareness of one's mental state while meditation is associated with an increasingly accurate meta-awareness. In the words of Semmens-Wheeler et al., (2012) hypnosis is "self-deception" while meditation is "self-insight". If self-deception and self-insight are on a continuum, i.e. mutually exclusive, these theories would predict that a hypnotic state and a meditative state cannot co-exist. If attempting to do so, they would either cancel each other out or only one would dominate, presumably after a rather conflicting battle of the two. In particular, since the active control condition consisted of a hypnotic induction followed by meditative suggestion, there should be a break or a disturbance of the hypnotic state at the time when the meditative suggestion began. This was not observed. One could say that patients shifted back and forth between hypnosis and meditation without showing so in overt behavior. We do not have data to assess this directly. It would require an assessment of the mental states of the patients while hypnotized/meditating and would be strengthened by having a condition with the same suggestions but without a formal hypnotic induction. Such an experiment would be well suited to test the merits of Cold Control theory in this instance, or at least it's account of the difference between hypnosis and meditation. For what it is worth, the hypnotist reported that there was no difference in the hypnotic depth between conditions as judged from observations and patients' own reports.

These considerations could be of theoretical importance, were the functioning of high-level cognition the only difference between patients with acquired brain injury and a healthy reference population. Naturally, this is not the case. All patients had traumatic experiences, many received disability pension, and some had physical disabilities. In these respects, and many more, they differ from the typical samples in hypnosis research.

activations are resolved. Maybe hypnotic suggestion are usually conflict-free so that the absence of conflict resolution does not cause problems.

3.3.4 Relevant hypnosis literature

There are at least 7 published single-case reports on hypnosis with brain injured patients. They sought to improve motivation for other therapy, improve compliance to other therapy, and improve impaired motor functions, e.g. in hemiplegia. These single-case studies were not pre-planned experiments but rather post-hoc reporting of interesting clinical cases. With a lack of pre-planned research questions and control groups, there are too many potential biases and unknowns to interpret their findings here. For the interested, here is the list:

- Manganiello et al. (1986), single case study.
- Holroyd et al. (1989), single case.
- Appel (1990), three cases.
- Vodovnik (1979), single case.
- Crasilneck & Hall (1970), three cases.
- Chappell (1961), two cases.
- Kirkner, Dorcus, and Seacat (1953), single case.

4 Discussion

We observed a large transfer effect in the hypnosis study contrary to the findings in the N-back study. I conclude that targeted hypnotic suggestion caused a generalized improvement of "untrained" behavior whereas computer-based training lead to very narrow improvements. I have argued that there was an effect of the hypnotic suggestions which could be uniquely attributed to the *meaning* of the targeted suggestions. In comparison, there was no inherently meaningful content in the computerized training tasks. I suggest that the gist of these results, and the results on the broader literature of cognitive enhancement and rehabilitation, may be summarized using three metaphors for the relationship between the intervention and the mind:

4.1.1 Metaphors for cognitive improvement

In the midst of many complex theories about human cognition, it can be useful to use intuition to arrive at a rough solution to a particular problem. I propose the following three metaphors as intuition pumps[37] on what kind effects to expect on high-level cognition from different intervention strategies.

Mind as a muscle, treatment as exercise: Computer training and paper-and-pencil tasks belong in this category. The idea is that by straining a cognitive function to the maximum of its capabilities "bottom-up", it will improve over time. The physical analogue is the muscle which is strengthened by repetitive exercise. Given the lack of transfer effects in the computer-training and the miniscule effects of Attention Process Training (SMD around

37 A term coined by the philosopher Daniel Dennett. An intuition pump is a thought experiment designed to arrive at a correct solution using intuition rather than bullet-proof reasoning. Effective as they are, they also have the potential to be misleading. Dennett originally used the term in a negative way to counter several thought experiments as misleading but later found them to be a useful way of conveying ideas, if used thoughtfully.

0.2) which were reviewed in the introduction, muscle-exercise interventions do not seem to be effective for higher-level cognition. However, the metaphor does not fail universally. It holds quite well for Constraint Induced Movement and Language Therapy, for example.

Mind as a plant, treatment as nourishment: Pharmaceuticals and physical exercise belong in this category. The idea is that by providing the right nutrition, the mind will regrow to its natural state. This is an analogue to nourishing a plant and letting it grow by itself. Given the small effect from 7 pharmaceutical and exercise studies reviewed in the introductory section (average SMD=0.18), the results are not encouraging for the plant-nourishment interventions targeting higher cognition.

Mind as a computer, treatment as programming: Suggestion-based interventions, such as hypnosis, meditation, and psychotherapy belong in this category. The idea is that the suggestion is a small piece of programming code which is interpreted by the subject and compiled into a more automated form. The programming metaphor fits well with the current evidence since we saw the semantics of suggestion be realized in behavioral changes. The non-targeted suggestion did as well, but was less efficient with respect to working memory behavior. Lastly, I reviewed three studies on Mindfulness Based Stress Reduction in the introduction, which yielded small effects relative to passive control groups (SMD=0.2). Merely providing "code" to the subject is not sufficient. The "code" has to be efficient and executed in the desired mental/physical context. Posthypnotic suggestion is often conditioned on a specific context, e.g. in the treatment of phobia. For example, Azulay et al. (2013) achieved large reductions in fatigue using MBSR and I have mentioned others who targeted anxiety using hypnotic suggestion (Sullivan et al., 1974; Fromm et al., 1964; Faymonville et al., 1997).

Other physical analogies could possibly capture the same intuitions. E.g. the computer metaphor could be equivalent to the mind as a top-down controlled organization. It is not the analogue that is important but the intuitions it carries.

4.1.2 Limitations

The limitations of the isolated conclusions for the three articles have been presented in their respective epilogues. With respect to the more general discussion, it is important to bear in mind that the N-back study and the hypnosis-study diverge on many parameters that does not pertain directly to the content of the treatments. I have compiled a list of factors that diverge between experiments but to which I argue that the difference in effects cannot be attributed (subject factors, location, duration, and therapist contact, see Appendix A). One factor, motivation, could confound the contrast between the studies to an unknown extent. Most subjects in the hypnosis study were highly motivated (see Figure 23) while motivation probably was lower for the subjects in the computer-training study. While we

did not assess this directly, the high dropout rate and the irregularity of training for subjects in the computer-training study is most likely an indicator of lower motivation.

Motivation is a confound for psychological theorizing but not for the evaluation of clinical utility. If subjects find hypnosis motivating and computer-training unmotivating, then that is a genuine property of the interventions as they are administered in real life and should therefore not be factored out as a nuisance parameter.

4.1.3 Clinical implications: revisiting cost-effectiveness

I introduced this thesis with a small review showing that current treatment approaches have low clinical value as judged from their cost-effectiveness. We are now in a position to add the N-back study, typical computer-based studies, and the hypnosis study to the table (see Table 13). Maximum-likelihood estimates of effect sizes are displayed to make it comparable to table 1. Effect sizes relative to the passive control group is included for the same reason.

In the N-back study, it took estimated average of around 60 minutes per patient to set up their user, instruct them, monitor their training, follow-up on missing training, and help with technical problems. Using the Hours Needed to Treat statistic, 31 hours of therapist time is needed for one patient to show a more favourable outcome than would be expected was he/she in the control group. In that time, however, 30 patients spent a total of 150 hours training in vain. For this reason, the N-back task is not recommended to improve behaviors in the domain of working memory. When applying the same analysis to a typical computer-based study in the review, one out of 70 patients would benefit from a genuine transfer effect. While that lucky patient would have trained for a median of 14.5 hours, the other 69 patients would collectively have trained 1000.5 hours in vain.

For the targeted hypnotic suggestions, it took an estimated 6.4 therapist to treat a subject relative to a passive control group. That is, for each patient that improves more than a passive control group, the therapist and other patients have spent only 2.4 hours in vain. The untargeted suggestion also seem feasible in this analysis with 12.8 Hours Needed to Treat and 8.8 therapist/patient hours spent in vain for each successful patient.

These estimates should be compared to the between 319 and 930 therapist hours required for Mindfulness Based Stress Reduction, Attention Process Training, and physical exercise. I conclude that hypnotic suggestion may be a viable new strategy for cognitive rehabilitation of patients with acquired brain injury in the chronic phase. Given the many ways in which this finding is novel and in some respects at odds with established knowledge, independent replication is needed. Whether hypnotic suggestion are effective in

the post-acute phase, for other patient groups, or even for healthy subjects, could also be fruitful topics for future studies.

Intervention	Target	BF > 3	Hour	SMD_c	NNT	HNT
N-back	Working memory	0 of 5	1.0	0.11	31.1	31.1
Visual search	Processing speed	0 of 4	5.0	-0.11	%	%
Targeted suggestions with active control	Working memory	2 of 2	4.0	0.84	3.3	13.2
Targeted suggestions with passive control	Working memory	2 of 2	4.0	1.72	1.6	6.4
Untargeted suggestions with passive control	Working memory	1 of 2	4.0	0.88	3.2	12.8
Naive effect of a typical computerized study		1.2 of 4	%	0.30	10.6	%
Transfer effect of a typical computerized study		1.2 of 4	%	0.05	70.0	%

Table 13: *Effect sizes, consistency of outcomes and economic feasibility of the interventions from the present thesis. BF: Bayes Factor. SMD_c: controlled effect size. NNT: Number Needed to Treat. HNT: Hours needed to treat. See Table 1 for further details.*

5 References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working Memory and Intelligence: The Same or Different Constructs? *Psychological Bulletin*, *131*(1), 30–60. http://doi.org/10.1037/0033-2909.131.1.30

Åkerlund, E., Esbjörnsson, E., Sunnerhagen, K. S., & Björkdahl, A. (2013). Can computerized working memory training improve impaired working memory, cognition and psychological health? *Brain Injury*, *27*(13-14), 1649–1657. http://doi.org/10.3109/02699052.2013.830195

Akinwuntan, A. E., De Weerdt, W., Feys, H., Pauwels, J., Baten, G., Arno, P., & Kiekens, C. (2005). Effect of simulator training on driving after stroke A randomized controlled trial. *Neurology*, *65*(6), 843–850.

Allen, S. N. (2008). *Stanford Hypnotic Susceptibility Scale, Form C: Norms for an American Indian sample*. WASHINGTON STATE UNIVERSITY. Retrieved from http://gradworks.umi.com/33/33/3333922.html

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*(1), 20–29. http://doi.org/10.1016/j.jecp.2009.11.003

Askay, S. W., Patterson, D. R., Jensen, M. P., & Sharar, S. R. (2007). A randomized controlled trial of hypnosis for burn wound care. *Rehabilitation Psychology*, *52*(3), 247–253. http://doi.org/10.1037/0090-5550.52.3.247

Azulay, J., Smart, C. M., Mott, T., & Cicerone, K. D. (2013). A Pilot Study Examining the Effect of Mindfulness-Based Stress Reduction on Symptoms of Chronic Mild Traumatic Brain Injury/Postconcussive Syndrome: *Journal of Head Trauma Rehabilitation*, *28*(4), 323–331. http://doi.org/10.1097/HTR.0b013e318250ebda

Baddeley, A. (2007). *Working Memory, Thought, and Action* (1st ed.). Oxford University

Press, USA.

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In Gordon H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. Volume 8, pp. 47–89). Academic Press.

Balk, E. M., Raman, G., Tatsioni, A., Chung, M., Lau, J., & Rosenberg, I. H. (2007). Vitamin B6, B12, and folic acid supplementation and cognitive function: a systematic review of randomized trials. *Archives of Internal Medicine*, *167*(1), 21–30.

Barber, T. X. (1965). "Hypnotic" Phenomena: A Critique of Experimental Methods. In J. E. Gordon (Ed.), *Handbook of Clinical and Experimental Hypnosis* (pp. 444–480). New York: Basic Books.

Barker-Collo, S. L., Feigin, V. L., Lawes, C. M. M., Parag, V., Senior, H., & Rodgers, A. (2009). Reducing Attention Deficits After Stroke Using Attention Process Training: A Randomized Controlled Trial. *Stroke*, *40*(10), 3293–3298. http://doi.org/10.1161/STROKEAHA.109.558239

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. http://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. Retrieved from http://CRAN.R-project.org/package=lme4

Bédard, M., Felteau, M., Gibbons, C., Klein, R., Mazmanian, D., Fedyk, K., & Mack, G. (2005). A mindfulness-based intervention to improve quality of life among individuals who sustained traumatic brain injuries: One year follow-up. *The Journal of Cognitive Rehabilitation*, *23*(1), 8–13.

Bédard, M., Felteau, M., Mazmanian, D., Fedyk, K., Klein, R., Richardson, J., … Minthorn-Biggs, M.-B. (2003). Pilot evaluation of a mindfulness-based intervention to improve quality of life among individuals who sustained traumatic brain injuries. *Disability & Rehabilitation*, *25*(13), 722–731. http://doi.org/10.1080/0963828031000090489

Benson, N., Hulac, D. M., & Kranzler, J. H. (2010). Independent examination of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV): What does the WAIS-IV measure? *Psychological Assessment*, *22*(1), 121–130. http://doi.org/10.1037/a0017767

Berger, J. O., & Sellke, T. (1987). Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence. *Journal of the American Statistical Association*, *82*(397), 112. http://doi.org/10.2307/2289131

Beugeling, N. (2015). *The effect of computerized online brain training on post-stroke fatigue, cognitive functioning, and daily living in stroke patients*. University of

Amsterdam.

Bongartz, W. (2000). Deutsche Normen für die Stanford Hypnotic Susceptibility Scale: Form C (SHSS: C). *Experimentelle Und Klinische Hypnose*, *16*(2), 123–133.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. http://doi.org/10.1037/0033-295X.108.3.624

Canivez, G. L., & Watkins, M. W. (2010). Investigation of the factor structure of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV): Exploratory and higher order factor analyses. *Psychological Assessment*, *22*(4), 827–836. http://doi.org/10.1037/a0020429

Carroll, L., Cassidy, J. D., Peloso, P., Borg, J., von Holst, H., Holm, L., … Pépin, M. (2004). Prognosis for mild traumatic brain injury: results of the who collaborating centre task force on mild traumatic brain injury. *Journal of Rehabilitation Medicine*, *36*(0), 84–105. http://doi.org/10.1080/16501960410023859

Chalmers, D. J. (2011). Frege's Puzzle and the Objects of Credence. *Mind*, fzr046.

Cha, Y.-J., & Kim, H. (2013). Effect of computer-based cognitive rehabilitation (CBCR) for people with stroke: a systematic review and meta-analysis. *NeuroRehabilitation*, *32*(2), 359–368.

Chen, S. H. A., Thomas, J. D., Glueckauf, R. L., & Bracy, O. L. (1997). The effectiveness of computer-assisted cognitive rehabilitation for persons with traumatic brain injury. *Brain Injury*, *11*(3), 197–210.

Chiesa, A., Calati, R., & Serretti, A. (2011). Does mindfulness training improve cognitive abilities? A systematic review of neuropsychological findings. *Clinical Psychology Review*, *31*(3), 449–464.

Chooi, W.-T., & Thompson, L. A. (2012). Working memory training does not improve intelligence in healthy young adults. *Intelligence*, *40*(6), 531–542. http://doi.org/10.1016/j.intell.2012.07.004

Chung, C. S., Pollock, A., Campbell, T., Durward, B. R., & Hagen, S. (2013). Cognitive rehabilitation for executive dysfunction in adults with stroke or other adult non-progressive acquired brain damage. In The Cochrane Collaboration (Ed.), *Cochrane Database of Systematic Reviews*. Chichester, UK: John Wiley & Sons, Ltd.

Cicerone, K. D. (2002). Remediation of "working attention" in mild traumatic brain injury. *Brain Injury*, *16*(3), 185–195. http://doi.org/10.1080/02699050110103959

Clark, A. (1989). *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*. Cambridge, MA, USA: MIT Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(03), 181–204. http://doi.org/10.1017/S0140525X12000477

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786.

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*(12), 547–552. http://doi.org/10.1016/j.tics.2003.10.005

Conway, A. R. ., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. . (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*(2), 163–183. http://doi.org/10.1016/S0160-2896(01)00096-4

Coulson, M. (2010). Confidence intervals permit, but don't guarantee, better inference than statistical significance testing. *Frontiers in Psychology*. http://doi.org/10.3389/fpsyg.2010.00026

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, *104*(2), 163–191. http://doi.org/10.1037/0033-2909.104.2.163

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(01), 87–185.

Cowan, N. (2005). *Working Memory Capacity*. New York: Psychology Press.

Cumming, G., & Finch, S. (2005). Inference by Eye: Confidence Intervals and How to Read Pictures of Data. *American Psychologist*, *60*(2), 170–180. http://doi.org/10.1037/0003-066X.60.2.170

Cumming, T. B., Marshall, R. S., & Lazar, R. M. (2013). Stroke, cognitive deficits, and rehabilitation: still an incomplete picture: Review. *International Journal of Stroke*, *8*(1), 38–45. http://doi.org/10.1111/j.1747-4949.2012.00972.x

Dade, L. A., Zatorre, R. J., Evans, A. C., & Jones-Gotman, M. (2001). Working Memory in Another Dimension: Functional Imaging of Human Olfactory Working Memory. *NeuroImage*, *14*(3), 650–660. http://doi.org/10.1006/nimg.2001.0868

Dahlin, E., Neely, A. S., Larsson, A., Backman, L., & Nyberg, L. (2008). Transfer of Learning After Updating Training Mediated by the Striatum. *Science*, *320*(5882), 1510–1512. http://doi.org/10.1126/science.1155466

Dahlin, E., Nyberg, L., Bäckman, L., & Neely, A. S. (2008). Plasticity of executive functioning in young and older adults: Immediate training gains, transfer, and long-term maintenance. *Psychology and Aging*, *23*(4), 720.

De Luca, R., Calabrò, R. S., Gervasi, G., De Salvo, S., Bonanno, L., Corallo, F., … Bramanti, P. (2014). Is computer-assisted training effective in improving rehabilitative outcomes after brain injury? A case-control hospital-based study. *Disability and Health Journal*, *7*(3), 356–360. http://doi.org/10.1016/j.dhjo.2014.04.003

Dennett, D. C. (1994). Cognitive science as reverse engineering several meanings of "Top-down" and "Bottom-up." *Studies in Logic and the Foundations of Mathematics*, *134*, 679–689. http://doi.org/10.1016/S0049-237X(06)80069-8

Derbyshire, S. W. G., Whalley, M. G., Stenger, V. A., & Oakley, D. A. (2004). Cerebral activation during hypnotically induced and imagined pain. *NeuroImage*, *23*(1), 392–401. http://doi.org/10.1016/j.neuroimage.2004.04.033

Dienes, Z., Lush, P., Semmens-Wheeler, R., Parkinson, J., Scott, R., & Naish, P. (n.d.). Hypnosis as self-deception; Meditation as self-insight. Retrieved from http://www.sussex.ac.uk/Users/rbs20/Dienes_et_al_Hypnosis_and_meditation_chapter(in_press).pdf

Dirette, D. K., Hinojosa, J., & Carnevale, G. J. (1999). Comparison of remedial and compensatory interventions for adults with acquired brain injuries. *The Journal of Head Trauma Rehabilitation*, *14*(6), 595–601.

Dou, Z. L., Man, D. W. K., Ou, H. N., Zheng, J. L., & Tam, S. F. (2006). Computerized errorless learning-based memory rehabilitation for Chinese patients with brain injury: A preliminary quasi-experimental clinical design study. *Brain Injury*, *20*(3), 219–225. http://doi.org/10.1080/02699050500488215

Dywan, J., & Bowers, K. (1983). The use of hypnosis to enhance recall. *Science*, *222*(4620), 184–185. http://doi.org/10.1126/science.6623071

Eddy, D. M. (1982). Probabilistic Reasoning in Clinical Medicine: Problems and Opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases* (pp. 249–267). Cambridge University Press.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193.

Engberg, A. W., & Teasdale, T. W. (2004). Psychosocial outcome following traumatic brain injury in adults: a long-term population-based follow-up. *Brain Injury*, *18*(6), 533–545. http://doi.org/10.1080/02699050310001645829

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach.

Journal of Experimental Psychology: General, *128*(3), 309–331. http://doi.org/10.1037/0096-3445.128.3.309

Ericsson, K. A. (2003). Exceptional memorizers: made, not born. *Trends in Cognitive Sciences*, *7*(6), 233–235.

Faymonville, M.-E., Mambourg, P. H., Joris, J., Vrijens, B., Fissette, J., Albert, A., & Lamy, M. (1997). Psychological approaches during conscious sedation. Hypnosis versus stress reducing strategies: a prospective randomized study. *Pain*, *73*(3), 361–367.

Feigin, V. L., Forouzanfar, M. H., Krishnamurthi, R., Mensah, G. A., Connor, M., Bennett, D. A., … others. (2014). Global and regional burden of stroke during 1990–2010: findings from the Global Burden of Disease Study 2010. *The Lancet*, *383*(9913), 245–255.

Feigin, V. L., Lawes, C. M., Bennett, D. A., & Anderson, C. S. (2003). Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century. *The Lancet Neurology*, *2*(1), 43–53.

Feinstein, A. R. (1998). P-values and confidence intervals: two sides of the same unsatisfactory coin. *Journal of Clinical Epidemiology*, *51*(4), 355–360.

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 69–78.

Fox, D. D., Lees-Haley, P. R., Earnest, K., & Dolezal-Wood, S. (1995). Base rates of postconcussive symptoms in health maintenance organization patients and controls. *Neuropsychology*, *9*(4), 606.

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not All Executive Functions Are Related to Intelligence. *Psychological Science*, *17*(2), 172–179. http://doi.org/10.1111/j.1467-9280.2006.01681.x

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. http://doi.org/10.1038/nrn2787

Fromm, E., Sawyer, J., & Rosenthal, V. (1964). Hypnotic simulation of organic brain damage. *The Journal of Abnormal and Social Psychology*, *69*(5), 482.

Furukawa, T. A., & Leucht, S. (2011). How to obtain NNT from Cohen's d: comparison of two methods. *PloS One*, *6*(4), e19070.

Gandhi, B., & Oakley, D. A. (2005). Does "hypnosis" by any other name smell as sweet? The efficacy of "hypnotic" inductions depends on the label "hypnosis." *Consciousness and Cognition*, *14*(2), 304–315. http://doi.org/10.1016/j.concog.2004.12.004

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ*, *292*(6522), 746–750.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition* (3 edition). Boca Raton: Chapman and Hall/CRC.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics: *Philosophy and the practice of Bayesian statistics*. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38. http://doi.org/10.1111/j.2044-8317.2011.02037.x

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review*, *102*(4), 684.

Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Blaha, M. J., … on behalf of the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. (2014). Heart Disease and Stroke Statistics--2014 Update: A Report From the American Heart Association. *Circulation*, *129*(3), e28–e292. http://doi.org/10.1161/01.cir.0000441139.02102.80

González, J. R. L., Valle-Inclán, F., & Díaz, A. A. (1996). Datos normativos de la Escala de Susceptibilidad Hipnótica de Stanford, Forma C, en una muestra española. *Psicothema*, *8*(2), 369–373.

Goodman, S. N., & Royall, R. (1988). Evidence and scientific research. *American Journal of Public Health*, *78*(12), 1568–1574.

Grant, N. I. E. (1980). Individual Differences in Working Memory and Reading. *Journal of Verbal Learning and Verbal Behavior*, 450.

Gray, J. M., Robertson, I., Pentland, B., & Anderson, S. (1992). Microcomputer-based attentional retraining after brain damage: A randomised group controlled trial. *Neuropsychological Rehabilitation*, *2*(2), 97–115. http://doi.org/10.1080/09602019208401399

Hadfield, J. D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software*, *33*(2), 1–22.

Hajek, V. E., Kates, M. H., Donnelly, R., & McGree, S. (1993). The effect of visuo-spatial training in patients with right hemisphere stroke. *Canadian Journal of Rehabilitation*, *6*(3), 175–186.

Halligan, P. W., & Oakley, D. A. (2013). Hypnosis and cognitive neuroscience: Bridging the gap. *Cortex*, *49*(2), 359–364. http://doi.org/10.1016/j.cortex.2012.12.002

Halpin, P. F., & Stam, H. J. (2006). Inductive Inference or Inductive Behavior: Fisher and Neyman: Pearson Approaches to Statistical Testing in Psychological Research (1940-1960). *The American Journal of Psychology*, *119*(4), 625. http://doi.org/10.2307/20445367

Hammond, D. C., Haskins-Bartsch, C., Grant Jr, C. W., & McGhee, M. (1988). Comparison of self-directed and tape-assisted self-hypnosis. *American Journal of Clinical Hypnosis*, *31*(2), 129–137.

Hampton, R. R. (2009). Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms? *Comparative Cognition & Behavior Reviews*, *4*, 17.

Han, C., & Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors. *Journal of the American Statistical Association*, *96*(455).

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working Memory Training May Increase Working Memory Capacity but Not Fluid Intelligence. *Psychological Science*, *24*(12), 2409–2419. http://doi.org/10.1177/0956797613492984

Henrich, J., Heine, S., & Norenzayan, A. (2008). The weirdest people in the world. *Unpublished Manuscript. University of British Columbia*.

Henson, R. (2005). What can functional neuroimaging tell the experimental psychologist? *The Quarterly Journal of Experimental Psychology Section A*, *58*(2), 193–233.

Higgins, J. P. T. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology*, *37*(5), 1158–1160. http://doi.org/10.1093/ije/dyn204

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *172*(1), 137–159. http://doi.org/10.1111/j.1467-985X.2008.00552.x

Hilgard, E. R. (1965). *Hypnotic susceptibility* (Vol. xiii). Oxford, England: Harcourt, Brace & World.

Hilgard, E. R. (1973). The domain of hypnosis: With some comments on alternative paradigms. *American Psychologist*, *28*(11), 972.

Hilgard, E. R. (1974). Toward a neodissociation theory: Multiple cognitive controls in human functioning. *Perspectives in Biology and Medicine*, *17*(3), 301–316.

Hilgard, E. R. (1977). *Divided consciousness: Multiple controls in human thought and action*. New York: Wiley-Interscience.

Hillary, F. G., Genova, H. M., Chiaravalloti, N. D., Rypma, B., & DeLuca, J. (2006). Prefrontal modulation of working memory performance in brain injury and disease. *Human Brain Mapping*, *27*(11), 837–847. http://doi.org/10.1002/hbm.20226

Holdnack, J. A., Zhou, X., Larrabee, G. J., Millis, S. R., & Salthouse, T. A. (2011).

Confirmatory factor analysis of the WAIS-IV/WMS-IV. *Assessment*, *18*(2), 178–191.

Huberty, C. J. (1993). Historical Origins of Statistical Testing Practices. *The Journal of Experimental Education*, *61*(4), 317–333. http://doi.org/10.1080/00220973.1993.10806593

Ishak, K. J., Platt, R. W., Joseph, L., Hanley, J. A., & Caro, J. J. (2007). Meta-analysis of longitudinal studies. *Clinical Trials*, *4*(5), 525–539. http://doi.org/10.1177/1740774507083567

Iverson, G. L. (2001). Interpreting change on the WAIS-III/WMS-III in clinical samples. *Archives of Clinical Neuropsychology*, *16*(2), 183–191.

Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J. M., Gavaghan, D. J., & McQuay, H. J. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials*, *17*(1), 1–12. http://doi.org/10.1016/0197-2456(95)00134-4

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, *105*(19), 6829–6833.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences*, *108*(25), 10081–10086. http://doi.org/10.1073/pnas.1103228108

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y. F., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning–implications for training and transfer. *Intelligence*, *38*, 625–635.

Johansson, B., Bjuhr, H., & Rönnbäck, L. (2012). Mindfulness-based stress reduction (MBSR) improves long-term mental fatigue after stroke or traumatic brain injury. *Brain Injury*, *26*(13-14), 1621–1628. http://doi.org/10.3109/02699052.2012.700082

Johnson, L. S., & Wiese, K. F. (1979). Live versus tape-recorded assessments of hypnotic responsiveness in pain-control patients. *International Journal of Clinical and Experimental Hypnosis*, *27*(2), 74–84. http://doi.org/10.1080/00207147908407548

Kane, M. J., Conway, A. R. ., Miura, T. K., & Colflesh, G. J. . (2007). Working memory, attention control, and the n-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *33*(3), 615.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working Memory Capacity and Fluid Intelligence Are Strongly Related Constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), 66–71. http://doi.org/10.1037/0033-2909.131.1.66

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The Generality of Working Memory Capacity: A Latent-Variable Approach to Verbal and Visuospatial Memory Span and Reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189–217. http://doi.org/10.1037/0096-3445.133.2.189

Kane, M. J., Poole, B. J., Tuholski, S. W., & Engle, R. W. (2006). Working memory capacity and the top-down control of visual search: Exploring the boundaries of "executive attention". *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 749–777. http://doi.org/10.1037/0278-7393.32.4.749

Kelly-Hayes, M., Beiser, A., Kase, C. S., Scaramucci, A., D'Agostino, R. B., & Wolf, P. A. (2003). The influence of gender and age on disability following ischemic stroke: the Framingham study. *Journal of Stroke and Cerebrovascular Diseases*, *12*(3), 119–126. http://doi.org/10.1016/S1052-3057(03)00042-9

Kerner, M. J., & Acker, M. (1985). Computer Delivery of Memory Retraining With Head Injured Patients. *Journal of Cognitive Rehabilitation*.

Kihlstrom, J. F. (2008). The domain of hypnosis, revisited. In M. R. Nash & A. J. Barnier (Eds.), *The Oxford handbook of hypnosis: Theory, research, and practice* (pp. 21–52). New York: Oxford University Press.

Kihlstrom, J. F. (2011). Prospects for de-automatization. *Consciousness and Cognition*, *20*(2), 332–334. http://doi.org/10.1016/j.concog.2010.03.004

Kihlstrom, J. F., Glisky, M. L., McGovern, S., Rapcsak, S. Z., & Mennemeier, M. S. (2013). Hypnosis in the right hemisphere. *Cortex*, *49*(2), 393–399. http://doi.org/10.1016/j.cortex.2012.04.018

Kim, Y. H., Ko, M. H., Seo, J. H., Park, S. H., Kim, K. S., Jang, E. H., … Cho, Y. J. (2003). Effect of Computer-Assisted Cognitive Rehabilitation Program for Attention Training in Brain Injury. *Journal of Korean Academy of Rehabilitation Medicine*, *27*(6), 830–839.

Kirsch, I., & Braffman, W. (2001). Imaginative Suggestibility and Hypnotizability. *Current Directions in Psychological Science*, *10*(2), 57–61. http://doi.org/10.1111/1467-8721.00115

Kirsch, I., Mazzoni, G., Roberts, K., Dienes, Z., Hallquist, M. N., Williams, J., & Lynn, S. J. (2008). Slipping into trance. *Contemporary Hypnosis*, *25*(3-4), 202–209. http://doi.org/10.1002/ch.361

Klatzky, R. L., & Erdelyi, M. H. (1985). The Response Criterion Problem in Tests of Hypnosis and Memory. *International Journal of Clinical and Experimental Hypnosis*, *33*(3), 246–257. http://doi.org/10.1080/00207148508406653

Klingberg, T. (2010). Training and plasticity of working memory. *Trends in Cognitive*

Sciences, *14*(7), 317–324. http://doi.org/10.1016/j.tics.2010.05.002

Kolmogorov, A. N. (1956). *Foundations of the Theory of Probability*. S.l.: Chelsea Pub Co.

Kosslyn, S. M., Thompson, W. L., Costantini-Ferrando, M. F., Alpert, N. M., & Spiegel, D. (2000). Hypnotic visual illusion alters color processing in the brain. *American Journal of Psychiatry*, *157*(8), 1279.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312.

Laidlaw, T. M. (1993). Hypnosis and Attention Deficits After Closed Head Injury. *International Journal of Clinical and Experimental Hypnosis*, *41*(2), 97–111. http://doi.org/10.1080/00207149308414541

Leclercq, M., & Sturm, W. (2002). Rehabilitation of attention disorders: a literature review. *Applied Neuropsychology of Attention. Theory, Diagnosis and Rehabilitation*, 341–264.

Lee, H., Kim, S.-W., Kim, J.-M., Shin, I.-S., Yang, S.-J., & Yoon, J.-S. (2005). Comparing effects of methylphenidate, sertraline and placebo on neuropsychiatric sequelae in patients with traumatic brain injury. *Human Psychopharmacology: Clinical and Experimental*, *20*(2), 97–104. http://doi.org/10.1002/hup.668

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lichtenberg, P., Shapira, H., Kalish, Y., & Abramowitz, E. G. (2009). Israeli Norms for the Stanford Hypnotic Susceptibility Scale, Form C. *International Journal of Clinical and Experimental Hypnosis*, *57*(2), 227–237. http://doi.org/10.1080/00207140802665492

Lifshitz, M., Aubert Bonn, N., Fischer, A., Kashem, I. F., & Raz, A. (2013). Using suggestion to modulate automatic processes: From Stroop to McGurk and beyond. *Cortex*, *49*(2), 463–473. http://doi.org/10.1016/j.cortex.2012.08.007

Lindeløv, J. K., Dall, J. O., Kristensen, C. D., & Aagesen, Marie Holt. (in review). Training and transfer effects of N-back training for brain injured and healthy subjects. *Neuropsychological Rehabilitation*.

Lin, Z. -c., Tao, J., Gao, Y. -l., Yin, D. -z., Chen, A. -z., & Chen, L. -d. (2014). Analysis of central mechanism of cognitive training on cognitive impairment after stroke: Resting-state functional magnetic resonance imaging study. *Journal of International Medical Research*, *42*(3), 659–668. http://doi.org/10.1177/0300060513505809

Liotti, M., Woldorff, M. G., Perez III, R., & Mayberg, H. S. (2000). An ERP study of the temporal course of the Stroop color-word interference effect. *Neuropsychologia*, *38*, 701–711.

Lodewyckx, T., Kim, W., Lee, M. D., Tuerlinckx, F., Kuppens, P., & Wagenmakers, E.-J.

(2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, *55*(5), 331–347. http://doi.org/10.1016/j.jmp.2011.06.001

Loetscher, T., & Lincoln, N. B. (2013). Cognitive rehabilitation for attention deficits following stroke. In The Cochrane Collaboration (Ed.), *Cochrane Database of Systematic Reviews*. Chichester, UK: John Wiley & Sons, Ltd.

Luaute, J., Halligan, P., Rode, G., Rossetti, Y., & Boisson, D. (2006). Visuo-spatial neglect: A systematic review of current interventions and their effectiveness. *Neuroscience & Biobehavioral Reviews*, *30*(7), 961–982. http://doi.org/10.1016/j.neubiorev.2006.03.001

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–280.

Lundqvist, A., Grundström, K., Samuelsson, K., & Rönnberg, J. (2010). Computerized training of working memory in a group of patients suffering from acquired brain injury. *Brain Injury*, *24*(10), 1173–1183. http://doi.org/10.3109/02699052.2010.498007

Malec, J., Jones, R., Rao, N., & Stubbs, K. (1984). Video game practice effects on sustained attention in patients with craniocerebral trauma. *Cognitive Rehabilitation*, *2*(4), 18–23.

Man, D. W., Soong, W. Y. L., Tam, S. F., & Hui-Chan, C. W. (2006). A randomized clinical trial study on the effectiveness of a tele-analogy-based problem-solving programme for people with acquired brain injury (ABI). *NeuroRehabilitation*, *21*(3), 205–217.

Ma, V. Y., Chan, L., & Carruthers, K. J. (2014). Incidence, Prevalence, Costs, and Impact on Disability of Common Conditions Requiring Rehabilitation in the United States: Stroke, Spinal Cord Injury, Traumatic Brain Injury, Multiple Sclerosis, Osteoarthritis, Rheumatoid Arthritis, Limb Loss, and Back Pain. *Archives of Physical Medicine and Rehabilitation*, *95*(5), 986–995.e1. http://doi.org/10.1016/j.apmr.2013.10.032

Mazer, B. L., Sofer, S., Korner-Bitensky, N., Gelinas, I., Hanley, J., & Wood-Dauphinee, S. (2003). Effectiveness of a visual attention retraining program on the driving performance of clients with stroke. *Archives of Physical Medicine and Rehabilitation*, *84*(4), 541–550. http://doi.org/10.1053/apmr.2003.50085

Mazzoni, G., Rotriquenz, E., Carvalho, C., Vannucci, M., Roberts, K., & Kirsch, I. (2009). Suggested visual hallucinations in and out of hypnosis. *Consciousness and Cognition*, *18*(2), 494–499. http://doi.org/10.1016/j.concog.2009.02.002

McArdle, J. J. (2009). Latent Variable Modeling of Differences and Changes with Longitudinal Data. *Annual Review of Psychology*, *60*(1), 577–605. http://doi.org/10.1146/annurev.psych.60.110707.163612

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 817.

McGurk, S. R., Twamley, E. W., Sitzer, D. I., McHugo, G. J., & Mueser, K. T. (2007). A meta-analysis of cognitive remediation in schizophrenia. *The American Journal of Psychiatry*, *164*(12), 1791–1802.

McMillan, T., Robertson, I. H., Brock, D., & Chorlton, L. (2002). Brief mindfulness training for attentional problems after traumatic brain injury: A randomised control treatment trial. *Neuropsychological Rehabilitation*, *12*(2), 117–125.

Melby-Lervåg, M., & Hulme, C. (2012). Is working memory training effective? A meta-analytic review. *Developmental Psychology*, *49*(2), 270–291. http://doi.org/10.1037/a0028228

Meythaler, J. M., Brunner, R. C., Johnson, A., & Novack, T. A. (2002). Amantadine to improve neurorecovery in traumatic brain injury–associated diffuse axonal injury: a pilot double-blind randomized trial. *The Journal of Head Trauma Rehabilitation*, *17*(4), 300–313.

Middleton, D. K., Lambert, M. J., & Seggar, L. B. (1991). Neuropsychological rehabilitation: microcomputer-assisted treatment of brain-injured adults. *Perceptual and Motor Skills*, *72*(2), 527–530.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Miller, M. E., & Bowers, K. S. (1993). Hypnotic analgesia: Dissociated experience or dissociated control?. *Journal of Abnormal Psychology*, *102*(1), 29–38.

Montgomery, G. H., Bovbjerg, D. H., Schnur, J. B., David, D., Goldfarb, A., Weltz, C. R., … Silverstein, J. H. (2007). A Randomized Clinical Trial of a Brief Hypnosis Intervention to Control Side Effects in Breast Surgery Patients. *JNCI Journal of the National Cancer Institute*, *99*(17), 1304–1312. http://doi.org/10.1093/jnci/djm106

Montgomery, G. H., David, D., Winkel, G., Silverstein, J. H., & Bovbjerg, D. H. (2002). The effectiveness of adjunctive hypnosis with surgical patients: a meta-analysis. *Anesthesia & Analgesia*, *94*(6), 1639–1645.

Moore, A., & Malinowski, P. (2009). Meditation, mindfulness and cognitive flexibility. *Consciousness and Cognition*, *18*(1), 176–186.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (n.d.). The Fallacy of Placing Confidence in Confidence Intervals. Retrieved from http://andrewgelman.com/wp-content/uploads/2014/09/fundamentalError.pdf

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2013). The humble Bayesian: Model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 68–75. http://doi.org/10.1111/j.2044-8317.2012.02067.x

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of Bayes Factors for Common Designs*. Retrieved from http://CRAN.R-project.org/package=BayesFactor

Morgan, C. L. (1894). *An introduction to comparative psychology*. London: W. Scott, limited.

Morrison, A. B., & Chein, J. M. (2010). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin & Review*, *18*(1), 46–60. http://doi.org/10.3758/s13423-010-0034-0

Moser, G., Trägner, S., Gajowniczek, E. E., Mikulits, A., Michalski, M., Kazemi-Shirazi, L., … others. (2013). Long-term success of GUT-directed group hypnosis for patients with refractory irritable bowel syndrome: a randomized controlled trial. *The American Journal of Gastroenterology*, *108*(4), 602–609.

Naäring, G. W. B., Roelofs, K., & Hoogduin, K. A. L. (2001). The stanford hypnotic susceptibility scale, form C: Normative data of a dutch student sample. *International Journal of Clinical and Experimental Hypnosis*, *49*(2), 139–145. http://doi.org/10.1080/00207140108410064

Neyman, J. (1957). "Inductive Behavior" as a Basic Concept of Philosophy of Science. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, *25*(1/3), 7. http://doi.org/10.2307/1401671

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241.

Nielsen, H., Knudsen, L., & Daugbjerg, O. (1989). Normative data for eight neuropsychological tests based on a Danish sample. *Scandinavian Journal of Psychology*, *30*(1), 37–45.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience*, *14*, 1105–1107. http://doi.org/10.1038/nn.2886

Norman, D. A., & Shallice, T. (1986). Attention to Action: Willed and Automatic Control of Behavior. In R. J. Davidson, G. E. Schwarts, & D. Shapiro (Eds.), *Consciousness and Self-regulation. Advances in Research and Theory* (pp. 1–18). New York: Plenum.

Nuys, D. V. (1973). Meditation, Attention, and hypnotic susceptibility: A correlational study. *International Journal of Clinical and Experimental Hypnosis*, *21*(2), 59–69. http://doi.org/10.1080/00207147308409306

Oakley, D. A., & Halligan, P. W. (2009). Hypnotic suggestion and cognitive neuroscience. *Trends in Cognitive Sciences*, *13*(6), 264–270.

Oakley, D. A., & Halligan, P. W. (2013). Hypnotic suggestion: opportunities for cognitive

neuroscience. *Nature Reviews Neuroscience*, *14*(8), 565–576. http://doi.org/10.1038/nrn3538

Park, N. W., & Ingles, J. L. (2001). Effectiveness of attention rehabilitation after an acquired brain injury: A meta-analysis. *Neuropsychology*, *15*(2), 199–210.

Park, N. W., Proulx, G.-B., & Towers, W. M. (1999). Evaluation of the Attention Process Training Programme. *Neuropsychological Rehabilitation: An International Journal*, *9*(2), 135. http://doi.org/10.1080/713755595

Pascalis, V. D., Bellusci, A., & Russo, P. M. (2000). Italian Norms for the Stanford Hypnotic Susceptibility Scale, Form C [1]. *International Journal of Clinical and Experimental Hypnosis*, *48*(3), 315–323. http://doi.org/10.1080/00207140008415249

Peirce, J. W. (2007). PsychoPy--Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13. http://doi.org/10.1016/j.jneumeth.2006.11.017

Perkins, D., & Salomon, G. (1989). Are Cognitive Skills Context-Bound? *Educational Researcher*, *18*(1), 16–25. http://doi.org/10.3102/0013189X018001016

Piccione, C., Hilgard, E. R., & Zimbardo, P. G. (1989). On the degree of stability of measured hypnotizability over a 25-year period. *Journal of Personality and Social Psychology*, *56*(2), 289–295.

Pinker. (1997). *How the Mind Works*. New York: W. W. Norton & Company.

Piskopos, M. (1991). *Neuropsychological changes following computer-driven cognitive remediation of severe traumatic closed head injured patients* (phd). Concordia University. Retrieved from http://spectrum.library.concordia.ca/3766/

Pitman, E. J. G. (1957). Statistics and Science. *Journal of the American Statistical Association*, *52*(279), 322. http://doi.org/10.2307/2280902

Ploughman, M., McCarthy, J., Bossé, M., Sullivan, H. J., & Corbett, D. (2008). Does Treadmill Exercise Improve Performance of Cognitive or Upper-Extremity Tasks in People With Chronic Stroke? A Randomized Cross-Over Trial. *Archives of Physical Medicine and Rehabilitation*, *89*(11), 2041–2047. http://doi.org/10.1016/j.apmr.2008.05.017

Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*, 59–63. http://doi.org/10.1016/j.tics.2005.12.004

Posada, D., & Buckley, T. (2004). Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. *Systematic Biology*, *53*(5), 793–808. http://doi.org/10.1080/10635150490522304

Poulin, V., Korner-Bitensky, N., Dawson, D. R., & Bherer, L. (2012). Efficacy of Executive

Function Interventions After Stroke: A Systematic Review. *Topics in Stroke Rehabilitation*, *19*(2), 158–171. http://doi.org/10.1310/tsr1902-158

Prokopenko, S. V., Mozheĭko, E. I., Levin, O. S., Koriagina, T. D., Chernykh, T. V., & Berezovskaia, M. A. (2012). Cognitive disorders and its correction in the acute period of ischemic stroke. *Zhurnal Nevrologii I Psikhiatrii Imeni S.S. Korsakova / Ministerstvo Zdravookhraneniia I Meditsinskoĭ Promyshlennosti Rossiĭskoĭ Federatsii, Vserossiĭskoe Obshchestvo Nevrologov [i] Vserossiĭskoe Obshchestvo Psikhiatrov*, *112*(8 Pt 2), 35–39.

Prokopenko, S. V., Mozheyko, E. Y., Petrova, M. M., Koryagina, T. D., Kaskaeva, D. S., Chernykh, T. V., … Bezdenezhnih, A. F. (2013). Correction of post-stroke cognitive impairments using computer programs. *Journal of the Neurological Sciences*, *325*(1-2), 148–153. http://doi.org/10.1016/j.jns.2012.12.024

Pulvermüller, F., Neininger, B., Elbert, T., Mohr, B., Rockstroh, B., Koebbel, P., & Taub, E. (2001). Constraint-induced therapy of chronic aphasia after stroke. *Stroke*, *32*(7), 1621–1626.

Quaney, B. M., Boyd, L. A., McDowd, J. M., Zahner, L. H., Jianghua He, Mayo, M. S., & Macko, R. F. (2009). Aerobic Exercise Improves Cognition and Motor Function Poststroke. *Neurorehabilitation and Neural Repair*, *23*(9), 879–885. http://doi.org/10.1177/1545968309338193

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*(3), 446.

Raz, A., & Campbell, N. K. J. (2011). Can suggestion obviate reading? Supplementing primary Stroop evidence with exploratory negative priming analyses. *Consciousness and Cognition*, *20*(2), 312–320. http://doi.org/10.1016/j.concog.2009.09.013

Raz, A., Shapiro, T., Fan, J., & Posner, M. I. (2002). Hypnotic Suggestion and the Modulation of Stroop Interference. *Archives of General Psychiatry*, *59*(12), 1155–1161. http://doi.org/10.1001/archpsyc.59.12.1155

Redick, T. S. (2015). Working memory training and interpreting interactions in intelligence interventions. *Intelligence*, *50*, 14–20. http://doi.org/10.1016/j.intell.2015.01.014

Redick, T. S., & Lindsey, D. R. B. (2013). Complex span and n-back measures of working memory: A meta-analysis. *Psychonomic Bulletin & Review*, *20*(6), 1102–1113. http://doi.org/10.3758/s13423-013-0453-9

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., … Engle, R. W. (2012). No Evidence of Intelligence Improvement After Working Memory Training: A Randomized, Placebo-Controlled Study. *Journal of Experimental Psychology: General*. http://doi.org/10.1037/a0029082

Repantis, D., Schlattmann, P., Laisney, O., & Heuser, I. (2010). Modafinil and methylphenidate for neuroenhancement in healthy individuals: A systematic review. *Pharmacological Research*, *62*(3), 187–206. http://doi.org/10.1016/j.phrs.2010.04.002

Riley, R. D. (2009). Multivariate meta-analysis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(4), 789–811.

Roark, J. B., Barabasz, A. F., Barabasz, M., & Lin-Roark, I. H. (2012). An Investigation of Taiwanese Norms for the Stanford Hypnotic Susceptibility Scale: Form C (Mandarin Chinese Translation). *International Journal of Clinical and Experimental Hypnosis*, *60*(2), 160–174. http://doi.org/10.1080/00207144.2012.648062

Roca, M., Parr, A., Thompson, R., Woolgar, A., Torralva, T., Antoun, N., … Duncan, J. (2010). Executive function and fluid intelligence after frontal lobe lesions. *Brain*, *133*(1), 234 –247. http://doi.org/10.1093/brain/awp269

Röhring, S., Kulke, H., Reulbach, U., Peetz, H., & Schupp, W. (2004). Effektivität eines neuropsychologischen Trainings von Aufmerksamkeitsfunktionen im teletherapeutischen Setting. *Neurol Rehab*, *10*, 239–246.

Rossetti, Y., Rode, G., Pisella, L., Farné, A., Li, L., Boisson, D., & Perenin, M.-T. (1998). Prism adaptation to a rightward optical deviation rehabilitates left hemispatial neglect. *Nature*, *395*(6698), 166–169.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes Factors for Model Selection in Regression. *Multivariate Behavioral Research*, *47*(6), 877–903. http://doi.org/10.1080/00273171.2012.734737

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. http://doi.org/10.3758/PBR.16.2.225

Ruff, R., Mahaffey, R., Engel, J., Farrow, C., Cox, D., & Karzmark, P. (1994). Efficacy study of THINKable in the attention and memory retraining of traumatically head-injured patients. *Brain Injury*, *8*(1), 3–14.

Russell, S. J. (2010). *Artificial intelligence: a modern approach* (3rd ed). Upper Saddle River: Prentice Hall.

SáNchez-ArmáSs, O., & Barabasz, A. F. (2005). Mexican Norms For The Stanford Hypnotic Susceptibility Scale, Form C. *International Journal of Clinical and Experimental Hypnosis*, *53*(3), 321–331. http://doi.org/10.1080/00207140590961448

SáNchez-Cubillo, I., PeriáñEz, J. A., Adrover-Roig, D., RodríGuez-SáNchez, J. M., RíOs-Lago, M., Tirapu, J., & Barceló, F. (2009). Construct validity of the Trail Making Test: Role of task-switching, working memory, inhibition/interference control, and

visuomotor abilities. *Journal of the International Neuropsychological Society*, *15*(03), 438. http://doi.org/10.1017/S1355617709090626

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2010). Hundred days of cognitive training enhance broad cognitive abilities in adulthood: Findings from the COGITO study. *Frontiers in Aging Neuroscience*, *2*.

Schmitter-Edgecombe, M. (2006). Implications of basic science research for brain injury rehabilitation: A focus on intact learning mechanisms. *The Journal of Head Trauma Rehabilitation*, *21*(2), 131–141.

Schneider, W. N., Drew-Cates, J., Wong, T. M., & Dombovy, M. L. (1999). Cognitive and behavioural efficacy of amantadine in acute traumatic brain injury: an initial double-blind placebo-controlled study. *Brain Injury*, *13*(11), 863–872.

Sedlmeier, P., Eberth, J., Schwarz, M., Zimmermann, D., Haarig, F., Jaeger, S., & Kunze, S. (2012). The psychological effects of meditation: A meta-analysis. *Psychological Bulletin*, *138*(6), 1139–1171. http://doi.org/10.1037/a0028168

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p Values for Testing Precise Null Hypotheses. *The American Statistician*, *55*(1), 62–71. http://doi.org/10.1198/000313001300339950

Semmens-Wheeler, R., & Dienes, Z. (2012). The contrasting role of higher order awareness in hypnosis and meditation. *The Journal of Mind–Body Regulation*, *2*(1), 43–57.

Serino, A., Ciaramelli, E., Di Santantonio, A., Malagù, S., Servadei, F., & Làdavas, E. (2006). Central executive system impairment in traumatic brain injury. *Brain Injury*, *20*, 23–32. http://doi.org/10.1080/02699050500309627

Shin, S. H., Ko, M. H., & Kim, Y. H. (2002). Effect of Computer-Assisted Cognitive Rehabilitation Program for Patients with Brain Injury. *Journal of Korean Academy of Rehabilitation Medicine*, *26*(1), 1–8.

Shipstead, Z., Hicks, K. L., & Engle, R. W. (2012). Cogmed working memory training: Does the evidence support the claims? *Journal of Applied Research in Memory and Cognition*, *1*(3), 185–193. http://doi.org/10.1016/j.jarmac.2012.06.003

Shipstead, Z., REdICk, T. S., & ENGlE, R. W. (2010). Does working memory training generalize? *Psychologica Belgica, 50*, *3*(4), 245–276.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*, *138*(4), 628–654. http://doi.org/10.1037/a0027473

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill*. Harvard University Press.

Smith, P. J., Blumenthal, J. A., Hoffman, B. M., Cooper, H., Strauman, T. A., Welsh-

Bohmer, K., … Sherwood, A. (2010). Aerobic Exercise and Neurocognitive Performance: A Meta-Analytic Review of Randomized Controlled Trials. *Psychosomatic Medicine*, *72*(3), 239–252. http://doi.org/10.1097/PSY.0b013e3181d14633

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs*, *74*(11), 1–29.

Stablum, F., Umiltà, C., Mogentale, C., Carlan, M., & Guerrini, C. (2000). Rehabilitation of executive deficits in closed head injury and anterior communicating artery aneurysm patients. *Psychological Research*, *63*(3-4), 265–278. http://doi.org/10.1007/s004269900002

Studenski, S., Duncan, P. W., Perera, S., Reker, D., Lai, S. M., & Richards, L. (2005). Daily Functioning and Quality of Life in a Randomized Controlled Trial of Therapeutic Exercise for Subacute Stroke Survivors. *Stroke*, *36*(8), 1764–1770. http://doi.org/10.1161/01.STR.0000174192.87887.70

Sturm, W. (2004). Functional reorganisation in patients with right hemisphere stroke after training of alertness: a longitudinal PET and fMRI study in eight cases. *Neuropsychologia*, *42*(4), 434–450. http://doi.org/10.1016/j.neuropsychologia.2003.09.001

Sturm, W., Dahmen, W., Hartje, W., & Willmes, K. (1983). Ergebnisse eines Trainingsprogramms zur Verbesserung der visuellen Auffassungsschnelligkeit und Konzentrationsfähigkeit bei Hirngeschädigten. *Archiv Für Psychiatrie Und Nervenkrankheiten*, *233*(1), 9–22.

Sturm, W., Fimm, B., Cantagallo, A., Cremel, N., North, P., North, P., … Leclercq, M. (2003). Specific Computerized Attention Training in Stroke and Traumatic Brain-Injured Patients. *Zeitschrift Für Neuropsychologie*, *14*(4), 283–292. http://doi.org/10.1024/1016-264X.14.4.283

Sturm, W., & Willmes, K. (1991). Efficacy of a reaction training on various attentional and cognitive functions in stroke patients. *Neuropsychological Rehabilitation*, *1*(4), 259–280. http://doi.org/10.1080/09602019108402258

Sturm, W., Willmes, K., Orgass, B., & Hartje, W. (1997). Do specific attention deficits need specific training? *Neuropsychological Rehabilitation*, *7*(2), 81–103.

Sullivan, D. S., Johnson, A., & Bratkovitch, J. (1974). Reduction of behavioral deficit in organic brain damage by use of hypnosis. *Journal of Clinical Psychology*, *30*(1), 96–98. http://doi.org/10.1002/1097-4679(197401)30:1<96::AID-JCLP2270300133>3.0.CO;2-A

Taub, E., Uswatte, G., Pidikiti, R., & others. (1999). Constraint-induced movement therapy:

a new family of techniques with broad application to physical rehabilitation-a clinical review. *Journal of Rehabilitation Research and Development*, *36*(3), 237–251.

Teasdale, T. W., Christensen, A. L., Willmes, K., Deloche, G., Braga, L., Stachowiak, F., … Leclercq, M. (1997). Subjective experience in brain injured patients and their close relatives: A European Brain Injury Questionnaire study. *Brain Injury*, *11*(8), 543–564.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. I. *Psychological Review*, *8*(3), 247.

Thut, G., & Pascual-Leone, A. (2010). A Review of Combined TMS-EEG Studies to Characterize Lasting Effects of Repetitive TMS and Assess Their Usefulness in Cognitive and Clinical Neuroscience. *Brain Topography*, *22*(4), 219–232. http://doi.org/10.1007/s10548-009-0115-4

Tournaki, N. (2003). The differential effects of teaching addition through strategy instruction versus drill and practice to students with and without learning disabilities. *Journal of Learning Disabilities*, *36*(5), 449–458.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505.

Vanderploeg, R. D., Crowell, T. A., & Curtiss, G. (2001). Verbal Learning and Memory Deficits in Traumatic Brain Injury: Encoding, Consolidation, and Retrieval. *Journal of Clinical and Experimental Neuropsychology (Neuropsychology, Development and Cognition: Section A)*, *23*(2), 185–195. http://doi.org/10.1076/jcen.23.2.185.1210

Vanhaudenhuyse, A., Boly, M., Balteau, E., Schnakers, C., Moonen, G., Luxen, A., … others. (2009). Pain and non-pain processing during hypnosis: A thulium-YAG event-related fMRI study. *Neuroimage*, *47*(3), 1047–1054.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. http://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189. http://doi.org/10.1016/j.cogpsych.2009.12.001

Wagstaff, M., Parkes, M., & Hanley, J. R. (2001). A comparison of posthypnotic amnesia and the simulation of amnesia through brain injury. *International Journal of Psychology and Psychological Therapy*, *1*(1), 67–78.

Weiand, C. (2006). *Neuropsychologische Behandlungsmethoden im Vergleich: eine randomisierte klinische Studie* (phd). Retrieved from http://kops.uni-konstanz.de/handle/123456789/10371

Weitzenhoffer, A. M., & Hilgard, E. R. (1962). *Stanford hypnotic susceptibility scale, Form C*. Palo Alto, CA: Consulting Psychologist Press.

Westerberg, H., Jacobaeus, H., Hirvikoski, T., Clevberger, P., Östensson, M.-L., Bartfai, A., & Klingberg, T. (2007). Computerized working memory training after stroke–A pilot study. *Brain Injury*, *21*(1), 21–29. http://doi.org/10.1080/02699050601148726

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests. *Perspectives on Psychological Science*, *6*(3), 291–298. http://doi.org/10.1177/1745691611406923

Whelan, R. (2010). Effective analysis of reaction time data. *The Psychological Record*, *58*(3), 9.

White, R. W. (1941). A preface to the theory of hypnotism. *The Journal of Abnormal and Social Psychology*, *36*(4), 477–505. http://doi.org/10.1037/h0053844

Wilson, B. A., Baddeley, A., Evans, J., & Shiel, A. (1994). Errorless learning in the rehabilitation of memory impaired people. *Neuropsychological Rehabilitation*, *4*(3), 307–326. http://doi.org/10.1080/09602019408401463

Wood, R. L., & Fussey, I. (1987). Computer-based cognitive retraining: a controlled study. *International Disability Studies*, *9*(4), 149–153.

Woody, E. Z., & Bowers, K. S. (1994). A Frontal Assault on Dissociated Control. In S. J. Lynn & J. W. Rhue (Eds.), *Dissociation: Clinical and theoretical perspectives* (pp. 52–79). New York: The Guilford Press.

Xu, X.-D., Ren, H.-Y., Prakash, R., Kumar, R., & Vijayadas. (2013). Outcomes of neuropsychological interventions of stroke. *Annals of Indian Academy of Neurology*, *16*(3), 319. http://doi.org/10.4103/0972-2327.116909

Yip, B. C., & Man, D. W. (2013). Virtual reality-based prospective memory training program for people with acquired brain injury. *Neurorehabilitation*, *32*(1), 103–115.

Zucchella, C., Capone, A., Codella, V., Vecchione, C., Buccino, G., Sandrini, G., … Bartolo, M. (2014). Assessing and restoring cognitive functions early after stroke. *Functional Neurology*, 1–8.

6 Appendices

6.1 Appendix A: What did not cause differential transfer in the N-back and hypnosis studies

Subject factors differed between studies**.** Age, gender, and chronicity did not affect outcomes of the hypnosis study, nor did any of these explain more variance than they added to the model in the review dataset. They are therefore unlikely to explain the major difference in outcome measures between the studies. For chronicity, however, the hypnosis study required subjects to be at least 1 year post injury and therefore data is lacking for effects of hypnosis in the post-acute phase.

Treatment location also differed between studies: Patients were hospitalized in the N-back paper whereas they were community-dwelling in the hypnosis study. It could be that the context somehow determined the outcome. This would be the claim that had the

participants in the hypnosis study instead been given computer-based training, they would have improved equally and the reverse: that had the hypnosis been given in a hospital, it would have been ineffective. This seems unlikely at face value. Furthermore, the review found no differential effect of training location on cognitive improvement.

Treatment duration: Patients trained on average 5 hours on the computer task whereas each phase of hypnosis treatment was approximately 4 hours of face-time, summing to 8 hours in total. Another aspect is intensity. Thus the duration of treatment cannot account for the difference. The computerized training was more regular with subjects training on average 3.5 days per week whereas there was one hypnosis sessions per week. It might be that one weekly 1-hour computer training session would be more optimal.

Therapist contact: the hypnosis was therapist-administered whereas experimenter contact was kept to a minimum in the N-back study. Therapist contact does not explain differences between hypnosis group A and B who had equal amount of therapist contact. It could be that therapist contact explains some of the effect in hypnosis group B relative to the passive controls and for both group A and B in phase 2, where there was no active control. However, the review found no effect of therapist contact so this is unlikely to have increased improvement on the transfer tasks in the N-back study. With respect to hypnosis, the only direct comparison of hypnotist-administered suggestions versus other media was conducted by Hammond et al. (1988) who administered suggestions using a hypnotist, audiotape, and through self-hypnosis with 48 subjects. The results showed a clear ranking so that subjective experience of the quality of hypnosis as well as objective performance was greatest with a hypnotist, lower with the audiotape and lowest for self-hypnosis. Means on scale 0-100 scale of subjective experience indicated that it was comparable between conditions (87, 79 and 70). No other actual effect sizes were reported. A meta-analysis on hypnosis for surgical patients found no statistically noticeable ($p < 5\%$) difference in the effect of 14 studies with hypnotist-administered suggestions (SMD=1.5) and 8 studies with audiotaped suggestions (SMD=0.55), although the absolute effect sizes seem to suggest a trend in favor of the former. Johnson and Weise (1979) found that subjects performed worse on an audiotaped hypnosis test compared to the same test administered by a hypnotist. The latter finding is less relevant for the present purposes since the effect did not covary with suggestibility.

Taken together, a minor part of the effect in the hypnosis study may be conditioned on a hypnotist being present, but the absence of a therapist is not thought to reduce effectiveness of the computer-based training. I conclude that this is not a major explanatory factor and that current evidence indicate that only minor decreases in effectiveness should be expected if the targeted suggestions were delivered using audio or video.

6.2 Appendix B: Statistical models

All statistics were done in R. I write the models here in lme4::lmer syntax although the random terms are specified differently across different packages. The following just serves to give the most crucial information but is, of course, not the full analysis script. The full analysis script is available upon request.

N-back paper

The N-back paper used a simple ANOVA-like mixed model, analyzed using lme4::lmer and BayesFactor::lmBF:

- **Time x group x treatment interaction, full and null:**
  dependent ~ time * group * treatment + (1|subject)
  dependent ~ time * group * treatment - time : group : treatment + (1|subject)
- **Within group interaction, full and null:**
  dependent ~ time * treatment + (1|subject)
  dependent ~ time + treatment + (1|subject)
- **The training data was modeled using the following power function:**
  power.f = deriv(~ k + a * x ^ b, namevec=c('k', 'a', 'b'), function.arg=c('x','k','a','b')).
- **Testing whether groups develop differently during training:**
  N ~ power.f(block_number, k, a, b) ~ (k|subject) + (a|treat) + (b|treat) + (a|group) + (b|group)
  N ~ power.f(block_number, k, a, b) ~ (k|subject) + (a|treat) + (b|treat)

JAGS was initially used for the power model but mixing was very poor.

Meta-analysis

The meta-analysis used the following, analyzed using metafor::rma.mv and MCMCglmm::MCMCglmm with known variances:

- **Naive model:**
  SMD_c ~ 1 + (1|contrast)
- **Testing e.g. whether SAME is different than SIMILAR:**
  SMD_c ~ 1 + b.TaU + b.passive + b.similar + b.same + (1|contrast)
  SMD_c ~ 1 + b.TaU + b.passive + b.similar + (1|contrast)
- **Testing whether the transfer effect (intercept in the model above) was explained by another covariate (full and null):**

$SMD_c \sim 1 + b.TaU + b.passive + b.similar + b.same + age + (1|contrast)$
$SMD_c \sim 1 + b.TaU + b.passive + b.similar + b.same + (1|contrast)$

Here the intercept correspond to "active + SAME" and the covariates are coded as:

- similarity: b.dissimilar < b.similar < b.same
- control: b.active < b.TaU < b.passive

… so e.g. all rows with b.same = 1 had b.similar = 1 and b.dissimilar = 1. That makes SAME the difference between SIMILAR and SAME.

For MCMCglmm, I used 100.000 samples and no thinning. MCMCglmm was used because mixing was very poor when implementing this model in JAGS with the product space method. The model was:

```
# Indicator variable for product space method
null.post ~ dbern(null.prior)
keep[1] <- 1
keep[2] <- null.post

# Hierarchical offsets per paper
sd.contrast ~ dunif(0, 2)  # variability between contrasts in SD
for(i in 1:n.id) {mean.contrast[i] ~ dnorm(0, sd.contrast^-2)}  # centered at zero

# Regression parameters
for(i in 1:5) {
  b[i] ~ dnorm(0, sd.prior^-2)
  b.keep[i] <- b[i] * keep[H[i]]  # product space indicator
}

# Fix to data
for(i in 1:length(y.obs)) {
  # underlying effect = covariates + random
  y[i] <- inprod(b.keep, X[i, ])  + mean.contrast[id[i]]

 # observed effect was generated by the observed SD and y
 y.obs[i] ~ dnorm(y[i], sd.obs[i]^-2)
}
```

Hypnosis paper

There was only one model and reduced versions of it. Modeled using lme4::lmer, BayesFactor::lmBF, and jags::jags:

- **Testing the targeted effect in phase 1 for TMT, full and null:**
  $TMT \sim 1 + C1 + Cb + C2 + BC1 + BCb + BC2 + ABC1 + ABCb + AC2 + (1|subject)$
  $TMT \sim 1 + C1 + Cb + C2 + BC1 + BCb + BC2 + ABCb + AC2 + (1|subject)$
- **Testing suggestibility on group B phase 2, full and null:**
  $TMT \sim 1 + C1 + Cb + C2 + BC1 + BCb + BC2*SHSS + ABC1 + ABCb + AC2 + (1|subject)$

TMT ~ 1 + C1 + Cb + C2 + BC1 + BCb + BC2 + ABC1 + ABCb + AC2 + (1| subject)

The effect was simultaneosly

- Cumulative over groups (control, group B, group A): C < B < A
- Cumulative over time (1st, break, 2nd): 1 < b < 2

Testing BC1 is testing that for phase 1, group B and the control experienced the same effect. Testing ABCb is testing that for the break, group A and B experienced the same effect.

JAGS model:

```
  # Priors on baseline population and subject
  # population mean for each group
  for(i in 1:n.group) {mean.pop[i] ~ dnorm(mean.pop.prior, sd.pop.prior^-2)}
  sd.pop ~ dunif(0, 100)  # flat prior on population SD.
  # subject-specific intercepts from group means and common SD
  for(i in 1:n.id) {subject[i] ~ dnorm(mean.pop[id.group[i]], sd.pop^-2)}

  # Model index for product space bayes factor estimation
  null.post ~ dbern(null.prior)
  keep[1] <- null.post  # test covariate
  keep[2] <- 1  # keep covariate

  # Priors on covariates and error
  sd.resid ~ dunif(0, 100)  # flat prior on residuals
  for(i in 1:n.b) {
    b.d[i] ~ dnorm(0, prior.scale.cov^-2)  # prior on regression parameters in stand.
units
    b.abs[i] <- b.d[i] * sd.pop  # in absolute units
    b[i] <- b.abs[i] * keep[effect.index[i]]  # add model selection. b is used in
regression
    b.abs.pred[i] ~ dnorm(b.abs[i], sd.resid^-2)  # posterior predictive for effects
    b.d.pred[i] <- b.abs.pred[i] / sd.pop
  }

  # Model and fix to observations
  for(i in 1:length(y.obs)) {
    y[i] <- subject[id[i]] + inprod(b, x[i,])  #model-based y.
    y.obs[i] ~ dnorm(y[i], sd.resid^-2)  # observe with uncertainty (error)
```

6.3 Appendix C: List of papers in review and meta-analysis

Åkerlund, E., Esbjörnsson, E., Sunnerhagen, K. S., & Björkdahl, A. (2013). Can computerized working memory training improve impaired working memory, cognition and psychological health? *Brain Injury*, *27*(13-14), 1649–1657. http://doi.org/10.3109/02699052.2013.830195

Beugeling, N. (2015). *The effect of computerized online brain training on post-stroke fatigue, cognitive functioning, and daily living in stroke patients*. University of Amsterdam.

Chen, S. H. A., Thomas, J. D., Glueckauf, R. L., & Bracy, O. L. (1997). The effectiveness of computer-assisted cognitive rehabilitation for persons with traumatic brain injury. *Brain Injury*, *11*(3), 197–210.

De Luca, R., Calabrò, R. S., Gervasi, G., De Salvo, S., Bonanno, L., Corallo, F., … Bramanti, P. (2014). Is computer-assisted training effective in improving rehabilitative outcomes after brain injury? A case-control hospital-based study. *Disability and Health Journal*, *7*(3), 356–360. http://doi.org/10.1016/j.dhjo.2014.04.003

Dou, Z. L., Man, D. W. K., Ou, H. N., Zheng, J. L., & Tam, S. F. (2006). Computerized errorless learning-based memory rehabilitation for Chinese patients with brain injury: A preliminary quasi-experimental clinical design study. *Brain Injury*, *20*(3), 219–225. http://doi.org/10.1080/02699050500488215

Gray, J. M., Robertson, I., Pentland, B., & Anderson, S. (1992). Microcomputer-based attentional retraining after brain damage: A randomised group controlled trial. *Neuropsychological Rehabilitation*, *2*(2), 97–115. http://doi.org/10.1080/09602019208401399

Kerner, M. J., & Acker, M. (1985). Computer Delivery of Memory Retraining With Head Injured Patients. *Journal of Cognitive Rehabilitation*.

Kim, Y. H., Ko, M. H., Seo, J. H., Park, S. H., Kim, K. S., Jang, E. H., … Cho, Y. J. (2003). Effect of Computer-Assisted Cognitive Rehabilitation Program for Attention Training in Brain Injury. *Journal of Korean Academy of Rehabilitation Medicine*, *27*(6), 830–839.

Lindeløv, J. K., Dall, J. O., Kristensen, C. D., & Aagesen, Marie Holt. (in review). Training and transfer effects of N-back training for brain injured and healthy subjects. *Neuropsychological Rehabilitation*.

Lin, Z. -c., Tao, J., Gao, Y. -l., Yin, D. -z., Chen, A. -z., & Chen, L. -d. (2014). Analysis of central mechanism of cognitive training on cognitive impairment after stroke: Resting-state functional magnetic resonance imaging study. *Journal of International Medical Research*, *42*(3), 659–668. http://doi.org/10.1177/0300060513505809

Lundqvist, A., Grundström, K., Samuelsson, K., & Rönnberg, J. (2010). Computerized training of working memory in a group of patients suffering from acquired brain injury. *Brain Injury*, *24*(10), 1173–1183. http://doi.org/10.3109/02699052.2010.498007

Malec, J., Jones, R., Rao, N., & Stubbs, K. (1984). Video game practice effects on sustained attention in patients with craniocerebral trauma. *Cognitive Rehabilitation*, *2*(4), 18–23.

Man, D. W., Soong, W. Y. L., Tam, S. F., & Hui-Chan, C. W. (2006). A randomized clinical trial study on the effectiveness of a tele-analogy-based problem-solving programme for people with acquired brain injury (ABI). *NeuroRehabilitation*, *21*(3), 205–217.

Middleton, D. K., Lambert, M. J., & Seggar, L. B. (1991). Neuropsychological rehabilitation: microcomputer-assisted treatment of brain-injured adults. *Perceptual and Motor Skills*, *72*(2), 527–530.

Piskopos, M. (1991). *Neuropsychological changes following computer-driven cognitive remediation of severe traumatic closed head injured patients* (phd). Concordia University. Retrieved from http://spectrum.library.concordia.ca/3766/

Prokopenko, S. V., Mozheĭko, E. I., Levin, O. S., Koriagina, T. D., Chernykh, T. V., & Berezovskaia, M. A. (2012). Cognitive disorders and its correction in the acute period of ischemic stroke. *Zhurnal Nevrologii I Psikhiatrii Imeni S.S. Korsakova / Ministerstvo Zdravookhraneniia I Meditsinskoĭ Promyshlennosti Rossiĭskoĭ Federatsii, Vserossiĭskoe Obshchestvo Nevrologov [i] Vserossiĭskoe Obshchestvo Psikhiatrov*, *112*(8 Pt 2), 35–39.

Prokopenko, S. V., Mozheyko, E. Y., Petrova, M. M., Koryagina, T. D., Kaskaeva, D. S., Chernykh, T. V., … Bezdenezhnih, A. F. (2013). Correction of post-stroke cognitive impairments using computer programs. *Journal of the Neurological Sciences*, *325*(1-2), 148–153. http://doi.org/10.1016/j.jns.2012.12.024

Röhring, S., Kulke, H., Reulbach, U., Peetz, H., & Schupp, W. (2004). Effektivität eines neuropsychologischen Trainings von Aufmerksamkeitsfunktionen im teletherapeutischen Setting. *Neurol Rehab*, *10*, 239–246.

Ruff, R., Mahaffey, R., Engel, J., Farrow, C., Cox, D., & Karzmark, P. (1994). Efficacy study of THINKable in the attention and memory retraining of traumatically head-injured patients. *Brain Injury*, *8*(1), 3–14.

Shin, S. H., Ko, M. H., & Kim, Y. H. (2002). Effect of Computer-Assisted Cognitive Rehabilitation Program for Patients with Brain Injury. *Journal of Korean Academy of*

Rehabilitation Medicine, *26*(1), 1–8.

Sturm, W. (2004). Functional reorganisation in patients with right hemisphere stroke after training of alertness: a longitudinal PET and fMRI study in eight cases. *Neuropsychologia*, *42*(4), 434–450. http://doi.org/10.1016/j.neuropsychologia.2003.09.001

Sturm, W., Dahmen, W., Hartje, W., & Willmes, K. (1983). Ergebnisse eines Trainingsprogramms zur Verbesserung der visuellen Auffassungsschnelligkeit und Konzentrationsfähigkeit bei Hirngeschädigten. *Archiv Für Psychiatrie Und Nervenkrankheiten*, *233*(1), 9–22.

Sturm, W., Fimm, B., Cantagallo, A., Cremel, N., North, P., North, P., … Leclercq, M. (2003). Specific Computerized Attention Training in Stroke and Traumatic Brain-Injured Patients. *Zeitschrift Für Neuropsychologie*, *14*(4), 283–292. http://doi.org/10.1024/1016-264X.14.4.283

Sturm, W., & Willmes, K. (1991). Efficacy of a reaction training on various attentional and cognitive functions in stroke patients. *Neuropsychological Rehabilitation*, *1*(4), 259–280. http://doi.org/10.1080/09602019108402258

Sturm, W., Willmes, K., Orgass, B., & Hartje, W. (1997). Do specific attention deficits need specific training? *Neuropsychological Rehabilitation*, *7*(2), 81–103.

Weiand, C. (2006). *Neuropsychologische Behandlungsmethoden im Vergleich: eine randomisierte klinische Studie* (phd). Retrieved from http://kops.uni-konstanz.de/handle/123456789/10371

Westerberg, H., Jacobaeus, H., Hirvikoski, T., Clevberger, P., Östensson, M.-L., Bartfai, A., & Klingberg, T. (2007). Computerized working memory training after stroke–A pilot study. *Brain Injury*, *21*(1), 21–29. http://doi.org/10.1080/02699050601148726

Wood, R. L., & Fussey, I. (1987). Computer-based cognitive retraining: a controlled study. *International Disability Studies*, *9*(4), 149–153.

Yip, B. C., & Man, D. W. (2013). Virtual reality-based prospective memory training program for people with acquired brain injury. *Neurorehabilitation*, *32*(1), 103–115.

Zucchella, C., Capone, A., Codella, V., Vecchione, C., Buccino, G., Sandrini, G., … Bartolo, M. (2014). Assessing and restoring cognitive functions early after stroke. *Functional Neurology*, 1–8.

SUMMARY

This thesis is an empirical investigation into two cost-effective treatment options for patients with acquired brain injury. Based on an experiment and a review, I argue that in general computer-based cognitive rehabilitation, as it is currently practiced, has virtually no effect on untrained tasks. That is, training does not cause cognitive transfer and thus does not constitute "brain training" or "brain exercise" of any clinical relevance.

A larger study found more promising results for a suggestion-based treatment in a hypnotic procedure. Patients improved to above population average in a matter of 4-8 hours, making this by far the most effective treatment compared to computer-based training, physical exercise, phamaceuticals, meditation, and attention process training.

The contrast between computer-based methods and the hypnotic suggestion treatment may be reflect a more general discrepancy between bottom-up and top-down processes although such a claim woukd require more empirical substantiation.

ISSN (online): 2246-123X
ISBN (online): 978-87-7112-290-9

AALBORG **UNIVERSITY PRESS**