This manuscript is published in:

Meiran, N., Dreisbach, G., & von Bastian, C. C. (2019). Mechanisms of working memory training: Insights from individual differences. *Intelligence, 73*, 78-87. doi: 10.1016/j.intell.2019.01.010

The final publication is available at https://doi.org/10.1016/j.intell.2019.01.010

Mechanisms of Working Memory Training: Insights from Individual Differences

Nachshon Meiran

Ben-Gurion University of the Negev, Beer-Sheva, Israel

Gesine Dreisbach

Regensburg University, Regensburg, Germany

Claudia von Bastian

University of Sheffield, Sheffield, UK

Authors' Note

This research was supported by DFG Grant DR 392/9-1 to GD and NM. Correspondence concerning this article should be either sent to Nachshon Meiran, nmeiran@bgu.ac.il, Gesine Dreisbach gesine.dreisbach@psychologie.uni-regensburg.de, or Claudia von Bastian c.c.vonbastian@sheffield.ac.uk

Abstract

Computerized working-memory training (WM), despite typically yielding large practice effects in the training task, transfers reliably almost only to similar tasks and barely transfers to Fluid Intelligence (Gf). We hypothesized that WM training tasks gradually become less related to Gf due to the development of task-specific skills that reduce reliance on WM. As a result, what is being trained in the advanced stages of training is weakly related to WM and Gf. This hypothesis leads to predicting that with training progression, there would be a gradual change in the rank-ordering of individuals (quasi-simplex) in the per-session scores of the training task coupled with a trend in reduction in Gf loadings of these scores. We reanalyzed individual differences in per-session scores in the training task from two moderately large-scale published studies. Results show that, as predicted, the correlations between per-session scores decreased with increasing temporal distance between sessions, suggesting a quasi-simplex pattern indicative of a gradual change in the rank-ordering of individuals. However, contrary to the prediction, the training tasks maintained or even tended to increase their Gf loading with training progression. We provide post-hoc accounts for these results, some which challenge prevalent assumptions beyond the attempt to improve Gf through improving WM.

Keywords: Working-memory, cognitive-enhancement, individual differences during training

Mechanisms of Working Memory Training: Insights from Individual Differences

1. Introduction

Working memory (WM), the ability to hold information available for complex cognition in the present moment (Oberauer & Hein, 2012) is relatively strongly correlated with fluid intelligence (Gf) (Shipstead, Harrison, & Engle, 2016). Therefore, it has been suggested that improving WM may enhance Gf. Indeed, improving WM has become a target of intense research efforts, with tools ranging from pharmacological interventions (Coghill et al., 2014) to meditation (Gallant, 2016), to video-games (Sala, Tatlidil, & Gobet, 2018). However, computerized training seems to remain as one of the most popular and most widely studied tools to improve WM (see von Bastian & Oberauer, 2014, for review).

Computerized WM training involves the repetitive execution of tasks such as N-back (Jaeggi, Buschkuehl, Jonides, & Perrig, 2008) or updating (Dahlin, Neely, Larsson, Bäckman, & Nyberg, 2008), typically over the course of multiple sessions. In an effort to maximize reliance on WM, many WM training interventions adaptively adjust task difficulty to individual performance so that trainees are constantly challenged (Klingberg, 2010). This approach rests on the assumption that a constant mismatch in task demand and ability would result in cognitive plasticity (Lövdén, Bäckman, Lindenberger, Schaefer, & Schmiedek, 2010; see von Bastian & Eschen, 2016, for evidence that it may be variability, not task difficulty, that challenges the flexibility of the cognitive system). For example, in WM training, task difficulty is typically raised by increasing the number of to-be-recalled items. As task difficulty is adjusted to individual performance, the stable increase in the level of difficulty with training progression is frequently used as indicator for training success.

The efficacy of WM training is usually assessed through transfer. It is assumed that the larger that functional overlap between the training task and an untrained task, the stronger the expected transfer of training benefits to the untrained task would be. The literature differentiates between *near transfer* seen in tasks tapping the same ability as the training task, and *far transfer* seen in gains in different yet related abilities. For WM training, benefits in untrained WM tasks are considered near transfer, and training-related benefits in Gf tests are considered far transfer.

Numerous meta-analyses summarize the large body of research on WM training and their conclusions are quite disappointing, with far-transfer benefits estimates ranging between complete absence (Melby-Lervåg, Redick, & Hulme, 2016; Sala et al., 2018), to very modest, perhaps almost negligible and short-lived (Au et al., 2014). Notably though, gains in the practiced tasks are consistently large, with trainees typically performing well above average by the end of training. For example, after 20 sessions of WM training, young adults can recall about eight items (e.g., von Bastian & Oberauer, 2013) – twice as many as the average cognitively healthy adult (Cowan, 2001). Even when controlling for baseline performance, task practice yields large gains with effect sizes (Cohen's *d*) ranging between 1 and 2

How can we explain the discrepancy between the absent or nearly absent far transfer effects on the one hand and the substantial improvements in the training tasks on the other hand? In this work, we examined the hypothesis (H1) that the *initially* strong correlation between Gf and WM training tasks declines in the course of training due to a gradual change in the makeup of abilities that contribute to WM training task success such as greater involvement in task-specific skills and strategies that are only weakly related to Gf. Hence, what is being trained in the advanced stages of WM training is, to a large extent, irrelevant to Gf. Confirming this

hypothesis would suggest that future research might find a way to keep the high Gf relatedness throughout the course of training resulting in improved training transferability.

Support for H1 comes from studies showing that performance may improve substantially through the development of task-specific skills and strategies that effectively reduce the WM demands of the training task. For example, the N-Back task requires determining whether the current item has been presented *n* trials beforehand. This decision *can* be made based on *controlled* recollection, that is, the actual retrieval of the item and its context, a process that relies on WM (Oberauer, 2005). However, it is also possible to answer correctly based solely on item familiarity by simply matching the probe against the memoranda whilst ignoring the context. This process does not involve (or barely involves) cognitive control (Jacoby, 1991) and, so, employing it reduces the WM demand of the N-Back task (Szmalec, Verbruggen, Vandierendonck, & Kemps, 2011). Thus, it is conceivable that training results in relatively greater or more efficient reliance on familiarity without improving recollection. Similarly, repeatedly practicing the same WM task could encourage participants to acquire task-specific strategies boosting performance in the training task without affecting the processes thought to be shared between WM and Gf. Indeed, there is now ample of evidence that trainees use strategies such as rehearsal or chunking during training that can lead to greater training task success (De Simoni & von Bastian, 2018; Dunning & Holmes, 2014; Laine, Fellman, Waris, & Nyman, 2018; Minear et al., 2016) but may arguably be weakly related to Gf.

H1 is in line with Rabbitt's (2004) notion that "tasks cease to be effective tests of executive function as soon as they are performed more than once" (p. 8). Along a similar line, recent theorizing (e.g., Kool & Botvinick, 2014) suggests that people generally try to avoid using cognitive effort as much as possible, and there is every reason to expect them to do so in WM

training tasks as well. Hence, when repeatedly practicing the same task for several weeks, it might actually encourage trainees to develop ways to reduce the need to use WM; for example by relying more strongly on familiarity-based processing (Szmalec et al., 2011) or by focusing on only a subset of memoranda (Atkinson, Baddeley, & Allen, 2017). Thus, the prediction following from H1 is of robust improvements in the training task which does not reflect improvement in the target ability but instead reflects the development of task-specific skills that reduce reliance on this ability. As a result, little if any far-transfer is expected. In fact, this is exactly what is frequently found.

The notion that the makeup of abilities that contribute to success on a task may gradually change with practice is also discussed in the literature on individual differences during skill acquisition. Two relevant issues from this literature are briefly reviewed below.

1.1. Quasi-simplex

The common finding when a task is administered multiple times (as do training tasks) is the quasi-simplex (Humphreys, 1960). The quasi-simplex refers to a specific pattern in the correlation between the per-session (and more generally, per-administration) scores of the training task. In this pattern, the highest correlations are seen between temporally adjacent sessions (e.g., between Test 1 and Test 2, Test 2 and Test 3, etc.) and these correlations decline with increasing temporal distance between sessions (e.g., the correlation between Test 1 and Test 3 is lower than that between Test 2 and Test 3, with the correlation between Test 1 and Test 5 being even lower than that). This finding shows that the rank-ordering of the individuals gradually changes. Specifically, for a high correlation between measures (e.g., Test 1 and Test 2) to be found, the rank ordering of individuals should be very similar in the two measures. According to the quasi-simplex, this is true for adjacent sessions where, for example, individuals

who were especially successful in Test 1 tend to be the same individuals who were especially successful in Test 2. However, since, according to the quasi-simplex, correlations become gradually lower with increasing temporal distance between sessions, this suggests that individuals who were especially successful in Test 1 would not necessarily be the same individuals who were successful in Test 8, for example. Note that, to our knowledge, the studies which examined quasi-simplex till now focused on tasks that do not tap on WM, especially the WM tasks that also involve adaptive difficulty.

1.2. Which abilities dictate success on the task at different stages of training?

The considerations above suggest a prediction of H1 that the impact that individual differences (as they were measured before training) have on task performance changes in the course of training. For example, if late stages of skill acquisition involve memory retrieval, it seems to follow that individual differences at these stages will be dominated by memory-retrieval ability. These issues have been examined in the skill-acquisition literature by testing the correlations between the abilities that were measured during pre-test and the per-session scores on the training task (for early studies, see Fleishman, 1972; Fleishman & Rich, 1963). According to Ackerman (1988), the initial stages of skill acquisition tax relatively general abilities such as declarative knowledge, spatial abilities, and notably: also working memory. Thus, initial performance on a task (and a cognitive test is typically a task) reflects general intelligence and similar broad abilities. With practice on the task, there is less and less reliance on the general abilities, and as a result, the correlation between the predictor (here, general abilities) and the outcome (here, performance on the training task) shows a declining pattern. A subsequent meta-analysis (Keil & Cortina, 2001) supported the particular prediction regarding declining validity of general abilities, but failed to support other predictions of Ackerman's

theory related to which abilities (e.g., perceptual speed) were predicted to show increased validities in more advanced stages of practice.

There are two possible interpretations of the declining validity effect (Alvares & Hulin, 1972; Dalal, Bhave, & Fiset, 2014). The *changing-task* interpretation is central to the prediction of H1, that is, that the makeup of abilities that contribute to success on the task gradually changes with practice. According to the *changing-person* interpretation, the test-takers change gradually with the passage of time or with training. Note that the changing-person perspective is not what H1 is about, but it is important to keep in mind the fact that declining validity has more than one interpretation. We note however that the changing-task perspective gained compelling support from a recent study (Dahlke, Kostal, Sackett, & Kuncel, 2018) in which predictors (cognitive test scores) and performance criteria (college performance) were measured over multiple occasions rather than just once, thus making it possible to disentangle the two interpretations of declining validity. This study showed that the validity changes were mostly due to the time at which the criterion was measured and were barely influenced by the time at which the predictor was assessed. These results strongly support the changing-task interpretation, at least in some contexts.

1.3. The present study

In the present work, we reanalyzed the results from two moderate size WM training studies (De Simoni & von Bastian, 2018; Shahar et al., 2018, henceforth DSVB and SH) to investigate the quasi-simplex and the correlations between Gf and performance on the training task as a function of the training session. We used these analyses (and interpretations thereof) to shed light on the reasons for the very limited transferability of WM training to Gf that we found. We reasoned that WM training would result in the development of task-specific skills without a

change in the underlying WM ability. These new skills that are unrelated to WM explain the performance improvement. Thus, the WM training task is predicted to be relatively strongly related to Gf, but only in the initial stages of practice before the new skills have been developed. The gradual change in the makeup of abilities that contribute to task success would thus be reflected in a quasi-simplex pattern (here, interpreted as reflecting a changing-task pattern) combined with a declining trend of correlations between Gf and the WM-training task.

The DSVB and SH data are optimally suited to test our predictions given that they comprised a moderately large number of sessions (but see Schmiedek, Lovden, & Lindenberger, 2010, for a training intervention with a much higher dosage), thus involving a sufficiently large number of per-session scores. Importantly, the number of participants per group was relatively high, as required for the assessment of correlations as opposed to mean level of performance. Although our focus was on the training tasks, we capitalized on the fact that both studies involved an active-control group who performed visual search (Redick et al., 2013). This feature enabled us to use the visual-search task as an important benchmark.

2. Method

Detailed methods are reported in the original studies. Both studies were approved by the relevant university ethics committees. Here, we summarize only the key characteristics of each study.

2.1. Participants

Table 1 lists the participant demographics from the two studies included in the present work. Volunteers were recruited to participate in a “cognitive training study” through the participant pool (DSVB) or among students who took a preparatory course (SH). Inclusion criteria were being at least highly proficient in German (DSVB)/Hebrew (SH), normal or

corrected-to-normal vision, no color blindness, no current psychiatric or neurological disorders, and no psychotropic drug use. Participants were randomly assigned to groups, following double-blind procedures (i.e., neither participants nor the experimenter administering the tests knew about the group allocation). The groups drawn from DSVB were compensated after study completion (CHF 120, approx. USD 120, or 10 course credits and CHF 20, approx. USD 20). In addition, participants received a bonus of up to CHF 50 (approx. USD 50) depending on the task level achieved by the end of training. The groups drawn from SH were also compensated after study completion (500 or 450 NIS, ~125-145$ USD).

Table 1

Demographic Data of Study Participants

Group	*n* included	Attrition *n* (withdrawn/excluded)	Age (*M*, *SD*)	Gender (f/m)
		DSVB		
Updating	59	22 (8/14)	22.61 (2.97)	40/19
Binding	66	11 (6/5)	24.55 (4.05)	45/21
Visual Search	72	3 (3/0)	23.81 (4.16)	49/23
		SH		
WM	72	1(1/0)	23.32	12/60
Visual Search	71	0	23.63	12/59

2.2. Cognitive Training Interventions

Training tasks are described in Table 2. Participants completed 20 training sessions (30-45 min each) within 5 weeks (DSVB) or 12-14 training sessions (30-45 min each, in two cohorts) within 8-10 weeks (SH). Training was adaptive in both studies, with the level of task difficulty being stepwise adjusted to individual performance. In DSVB, training was self-administered at home using Tatool (von Bastian, Locher, & Ruflin, 2013), www.tatool.ch). After each session, training data were automatically uploaded to a webserver, allowing for constantly monitoring participants' training performance and compliance. In SH, training was programmed with E-Prime and training sessions took place in a designated classroom located at Ben-Gurion University campus. The class contained 14 testing positions, each comprising of a desk and a desktop computer.

Table 2

Training Tasks

Group/Task	Description
	DSVB
Updating	
Digits	Memorize a set of digits and update by applying simple arithmetic operations to them (Lewandowsky, Oberauer, Yang, & Ecker, 2010).
Letters	Memorize a set of letters and update by mentally shifting them up to three positions forward or backward in the alphabet (Lewandowsky et al., 2010).
Arrows	Memorize a set of arrows and update by rotating them for 45 degrees clockwise or counterclockwise (Harrison et al., 2013).

Locations	Memorize the locations of a set of circles in a grid and update by mentally shifting them to an adjacent cell as indicated by an arrow (cf. Lewandowsky et al., 2010).
Binding	
Noun-verb	Memorize a series of associations between nouns and verbs (Wilhelm, Hildebrandt, & Oberauer, 2013).
Symbol-digit	Memorize a series of associations between mathematical symbols and digits (cf. Wilhelm et al., 2013).
Fractal-location	Memorize a series of associations between fractals and their location in a row of boxes on the grid (Oberauer, 2005).
Color-location	Memorize a series of associations between colored circles and their locations in a 4 x 4 grid (cf. Oberauer, 2005).
Visual Search	
Numbers	Search for a “3” among horizontally and vertically presented “8”s (Kane, Poole, Tuholski, & Engle, 2006).
Letters	Search for a “T” among horizontally and vertically presented “I”s (cf. Harrison et al., 2013).
Arrows	Search for a single-headed arrow among double-headed arrows (cf. Kane et al., 2006).
Circles	Search for a circle with one gap among circles with two gaps (von Bastian, Langer, Jäncke, & Oberauer, 2013).

Procedural-WM	Combination of the Task-switching and N-back paradigm where participants switched between location and object classification, while reacting according to information presented in the Nth trial (Nitzan Shahar & Meiran, 2015)
Visual Search	Report whether a target letter ('F') placed among array of distractors is facing right or left (Redick et al., 2013).

2.3. Cognitive Ability Tests

To evaluate training effectiveness, participants completed 28 (DSVB) or 6 tasks (SH) before and after training. Table 3 lists descriptions and details of the tasks included in the present work.

Table 3

Cognitive Ability Tests

Group/Task	Description	#Trials Time allowed
	DSVB	
Diagramming Relationships	Determine the semantic relationship between three nouns and, out of five options, identify the Venn diagram that represents it best (Ekstrom, French, Harman, & Derman, 1976).	30 8 min
Letter Sets	Determine the logical pattern underlying a series of letter sets and identify the one set that deviates from the four others (Ekstrom et al., 1976).	30 14 min

Locations Test	Determine the logical pattern underlying the spatial distribution of crosses spread across rows of dashes and select the correct location for placing the next cross out of five options (Ekstrom et al. 1976).	28 12 min
RAPM	Complete a pattern by choosing the correct piece out of eight options (Arthur & Day, 1994).	28 15 min
	SH	
Operation Span	Memorizing letters while solving simple math equations (Unsworth, Heitz, Schrock, & Engle, 2005).	15 (3 of set size 3-7)
Symmetry Span	Memorizing spatial locations while performing a symmetry judgement (Unsworth et al., 2005).	12 (3 of set-size 2-5)
Comprehension	Follow complex instructions (e.g., "In the following digit sequence, count how many times does the digit 7 appear after an even number"), (Fischman, 1982).	20 7 min
Locations	Rule finding test - X mark is appearing in several location and participant should indicate the next X position according to a certain rule. (Ekstrom et al., 1976)	14 6 min
Choice RT	6-choice RT tasks performed on letters, digits, or shapes using arbitrary or non-arbitrary stimulus-response mapping (Shahar, Teodorescu, Usher, Pereg, & Meiran, 2014)	100 trials each

NEXT-task	A task made on miniblocks. In each, a novel set of instructions for a 2-choice task is presented. Afterwards, stimuli are presented in red color, and participants advance the screen by pressing a predetermined key. When the stimuli appear in green (twice only), participants apply the novel instructions (Meiran, Pereg, Kessler, Cole, & Braver, 2015)	55 miniblocks

We begin by examining the correlations between Gf, as assessed during pretest, and per-session scores reflecting success in the training task.

3. Correlations between Gf and success in the training task

3.1. Analyses

The correlations were assessed in two steps. First, using the full sample from each study, we assessed Gf by using Bayesian confirmatory factor analysis (BCFA). We then estimated the factor loadings of the training tasks per session using a BCFA variant of extension analysis (Dwyer, 1937). This method was required for two reasons. First, including the per-session scores in the pretest BCFA would have changed the meaning of Gf factors by confounding baseline ability with person/task related changes. Second, the extension analysis enabled us to use the entire sample within each study for the BCFA and not only the relatively smallish (for Structural Equations Modeling) sample of each experimental group. Hence, we could obtain a trustworthy factor solution with the relatively larger *N*s from each study. As each group completed different training tasks, we estimated the factor loadings of the session scores separately for each group, where the *N*s were considerably smaller. The fact that in each one of these within-group per-

session analyses we only estimated the Session loading partly compensates for the inaccuracy associated with estimation based on a small *N*. We report a more standard least-squares based extension analysis in Supplementary Materials online. The qualitative conclusions were similar in the two sets of analyses.

In all analyses, we employed the maximal number of participants with non-missing data (missing data were rare to begin with). All analyses were performed with R (R Core Team, 2014). We used the 'blavaan' package (Merkle & Rosseel, 2018) for the present analyses and used the default priors. We used the 'psych' package (Revelle, 2017) for PCA using the principal() function, and the 'bayesboot' package (Bååth, 2016) for Bayesian bootstrapping to determine the 95% credible intervals.

3.2. Results

In SH, we estimated Gf using four reasoning tests, two from the pretest and two PET scores (Nevo & Oren, 1986, the Israeli SAT, serving as entrance scores for the University) that were provided by the participants. In DSVB we used the four reasoning tests listed in Table 3. In both cases, we fitted three measurement models including M1: a single factor solution (tentatively labelled Gf); M2: a two correlated factor solution in which the factors were verbal and spatial; and M3: a single factor solution, in which the errors of the verbal/spatial tasks were correlated. The three models were evaluated using Bayesian Information Criterion (BIC, Schwarz, 1978). (Please keep in mind that we were mostly interested in finding a relatively reasonable assessment of Gf.) In DSVB, the BIC values were 2140.741, 2147.828, and 2149.896, for M1 through M3, respectively. These results favor the single factor solution (M1). In SH, the BIC values were 1961.171, 1958.330, and 1964.131, respectively. Although M1 was slightly inferior relative to M2, the difference in BIC corresponds to a Bayes Factor of 4.139 (see

Neath & Cavanaugh, 2012, regarding the computation). Given that (a) were interested in estimating Gf, and (b) that Gf has been found to be nearly perfectly correlated with g', and (c) that M2 implies a hierarchical structure (with g' at the top of the hierarchy) that could not be estimated here given the small number of indicators, we adopted M1 as the solution. Table 4 lists the factor pattern found in the two studies. The 95% Highest Posterior Density interval excluded zero for all loadings (meaning that the hypothesis concerning zero loading is not supported by the data). Given that the tests all involved a significant reasoning aspect, we tentatively describe the single factor as Gf. We acknowledge the fact that, given the low loadings of ETS-Locations and Comprehension in SH, the single factor in that study is not an ideal estimate of the Gf construct, and it may be slightly biased towards PET, seemingly relying on a combination of problem-solving and crystalized abilities.

Table 4

Factor Pattern (Standardised loadings)

Task	Gf loading	Lower HPD	Upper HPD
	DSVB		
Relations	.694	.545	.855
Locations	.549	.394	.706
RAPM	.616	.463	.769
Letters	.718	.564	.872
	SH		
Comprehension	.369	.161	.579
Locations	.224	.034	.427

PET-V	.716	.482	.968
PET-Q	.617	.401	.842

Note. The factor pattern determined with the full sample of each study. Highest Posterior Density (HPD) interval is between 2.5% (Lower) and .975 (higher).

Next, we evaluated how training performance loaded on the single factor. To reduce the very few missing values in DSVB (6 in Binding and one in VS), we computed session scores by averaging performance across the four training tasks that were included in each session in DSVB. Evaluation was based on series of BCFAs. In each BCFA, there were five variables including the four variables used to define Gf, for which the loadings were fixed to those in Table 4, with the Session score's loading being a free parameter. The results are presented in Figures 1 and 2.

As expected, Gf loadings of WM tasks were generally higher than those of visual search tasks which would be expected. In fact, this is actually why the visual search task has been used as the active control task (Redick et al., 2013). Importantly, in none of the five groups there was a systematic decline in loadings with training progression and the trends were either stable or even tended upward. The only visible exception is the visual-search group in SH where the loadings initially declined but then returned to their original level.

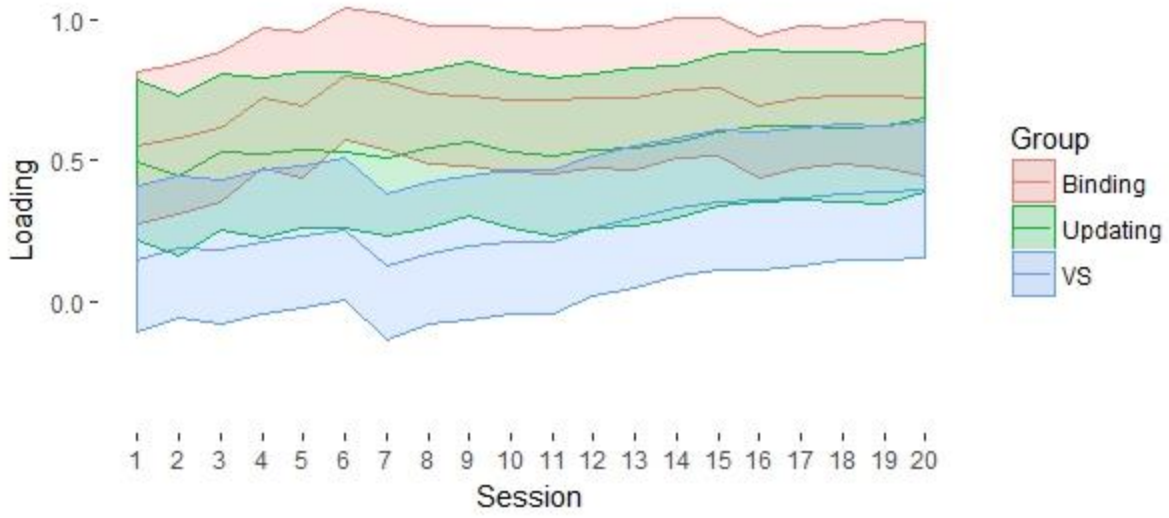

Figure 1. Session Score Loadings on Gf – DSVB. Bands indicate 95% Highest Posterior Density Intervals. VS = visual search.

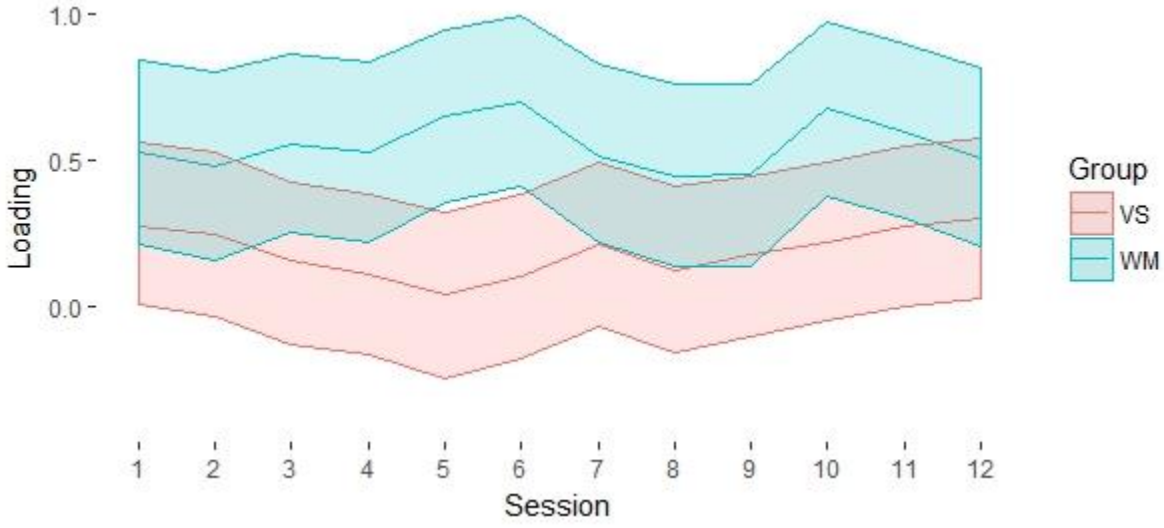

Figure 2. Session Score Loadings on Gf – SH: *Bands indicate Highest Posterior Density Intervals. VS = visual search; WM = procedural working memory.*

Since the loadings of observed variables might be influenced by reliability, we decided to estimate the reliability of the Session scores. For this purpose, we correlated a given Session

score with that of the next Session score, with the correlation representing retest reliability, yielding *k-1* reliability estimates for each study (*k* being the number of sessions). Figures 3 and 4 present these estimates (credible intervals were assessed using Bayesian bootstrapping with 1,000 samples). As can be seen, the reliability of Session scores was nearly perfect in DSVB while in SH it was moderate. Additionally, in the visual-search group in SH, it showed a slight upward trend especially across the initial sessions. Given the fact that the loadings of that group showed the opposite pattern in that range, this pattern of loadings is not easily explained as reflecting a change in reliability. To summarize this analysis, the results show that the trends of the Gf loadings probably do not reflect changing reliability.

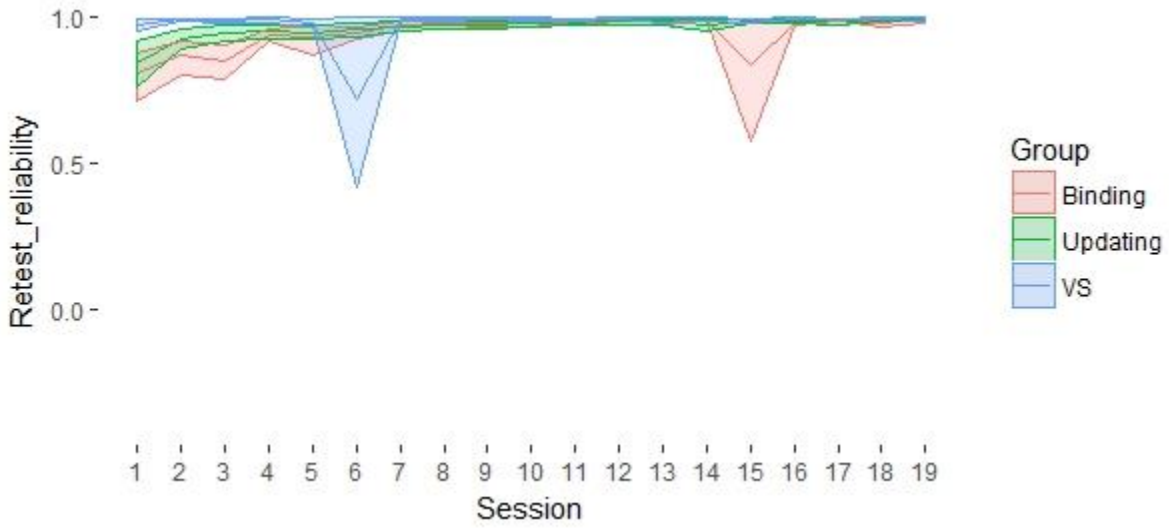

Figure 3. Retest Reliability of Session Scores – DSVB: Each Reliability Estimate is the Pearson Correlation between the Given Session's score and that of the Next Session. Bands represent 95% Credible Intervals estimated using Bayesian bootstrapping with 1,000 samples. VS = visual search.

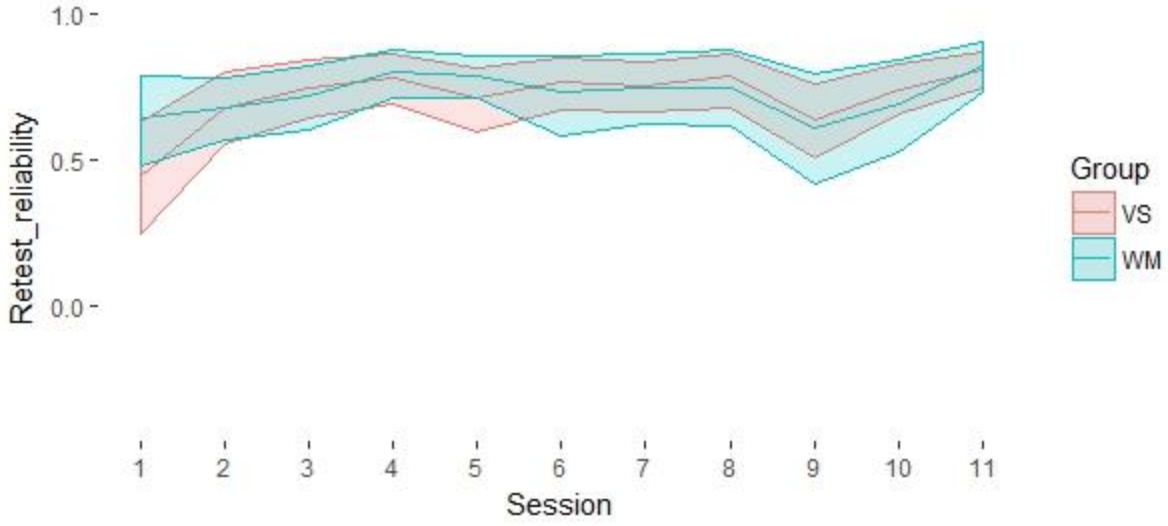

Figure 4. Retest Reliability of Session Scores – SH: Each Reliability Estimate is the Pearson Correlation between the Given Session's score and that of the Next Session. Bands represent 95% Credible Intervals estimated using Bayesian bootstrapping with 1,000 samples. VS = visual search; WM = procedural working memory.

To gain insight as to why the upward trend in Gf loadings occurred in DSVB, we used a novel method that capitalized on the fact that each session in DSVB involved four different tasks. Specifically, we assessed the similarity of the rank-ordering of the individual differences across the four tasks of a given session. To this end, we conducted a PCA on the four scores and examined the eigenvalue of the first unrotated component, which reflects the proportion of shared variance across the four scores. Given that there were four tasks, the maximal possible eigenvalue was 4. As illustrated in Figure 5, the results indicate that the individual differences in the four training tasks of each group became increasingly similar to one another with training progression, with non-overlapping credible intervals between the first and last session indicating a relatively clear-cut result.

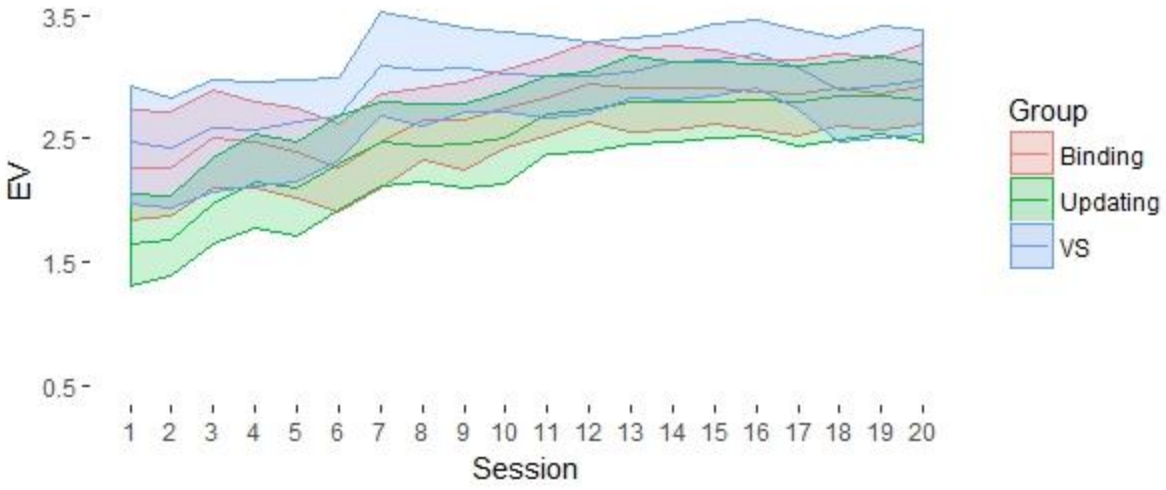

Figure 5. The degree of similarity of individual differences (Eigenvalue, EV, maximal value is 4) across the four training tasks in DSVB as a function of Session. Bands indicate 95% credible intervals estimated using Bayesian bootstrapping with 1,000 samples. VS = visual search.

3.3. Summary

To summarize, despite differences between studies, one conclusion emerges: We did *not* observe a pattern of decreasing Gf loadings of the WM training tasks. Instead, we either found stability (SH) or even an increasing trend (DSVB). Actually, there was no systematic decrease in loadings even in visual-search training. It is unlikely that these trends emerged as a result of a change in the reliability of the Session scores, given the near perfect reliability in DSVB and the opposite-to-predicted trend in SH. Last, we were able to show that the individual differences in the four training tasks used in each group in DSVB became increasingly similar over the course of training.

Our next section addresses the quasi-simplex pattern of correlations.

4. Quasi-Simplex

4.1.Analyses

The quasi-simplex was evaluated by computing the mean correlation (through Fisher's *Z* transformation, using the relevant functions in the 'psych' package, Revelle, 2017) between session scores as a function of the temporal distance between the sessions.

4.2. Results

Figures 6 and 7 visualize the quasi-simplex. The maximal possible distance between sessions was 19 (between Sessions 1 and 20) in DSVB and 11 (between Session 1 and Session 12, the last session in one of the cohorts in that study) in SH. In this analysis, more correlations are averaged for the short distances than for the long distances, of course (e.g., in SH, for Distance = 1 these were the correlations between Sessions 1-2, Session 2-3 etc., i.e., 11 such pairs; however, for Distance = 11 it was only one correlation). This should not systematically change the size of the correlations but should influence the stability of the estimates. Indeed, the credible intervals are narrower for the shorter temporal distances. As before, credible intervals were assessed by using Bayesian boostrapping with 1,000 samples.

Most importantly, the results from all groups in both studies show a clear reduction in correlations between session scores with increasing temporal distance between the sessions. This is seen in the fact that the mean correlation in the last sessions fall clearly outside the 95% credible interval of the first sessions and vice versa. Note that the fact that the quasi-simplex replicates across studies is not influenced by the difference in the number of training sessions. Specifically, in DSVB, the mean correlation at Distance = 11 (the maximal distance in SH) and Distance = 1 falls outside the credible interval of one another, thus supporting the hypothesis that the mean correlation has changed. In summary, the results clearly indicate a quasi-simplex pattern in the correlations matrices.

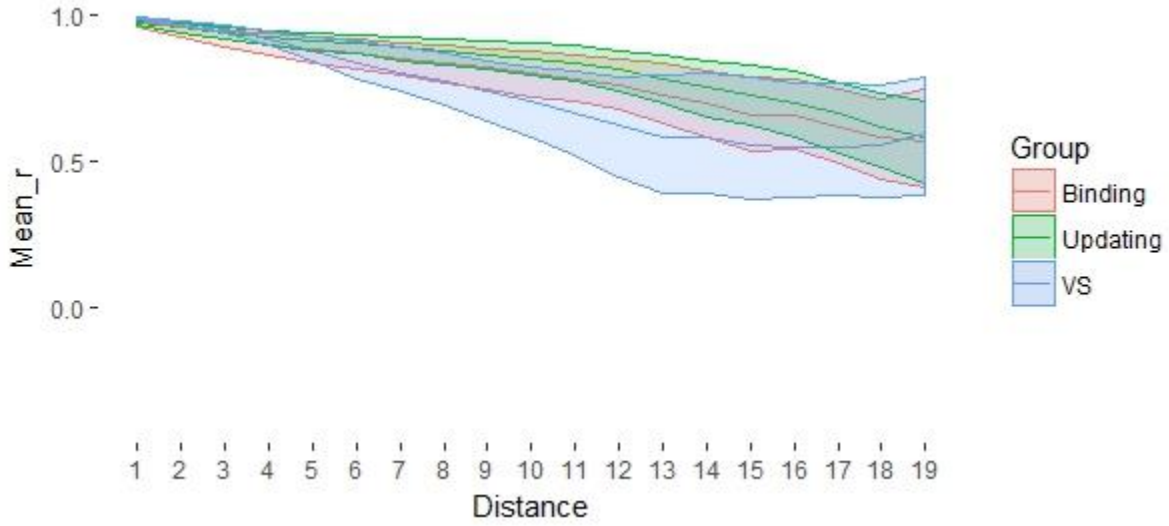

Figure 6. Quasi-simplex pattern - DSVB. Mean Pearson correlation between session scores (Mean_r) as a Function of the temporal distance between sessions. Bands indicate 95% credible intervals estimated using Bayesian bootstrapping with 1,000 samples. VS = visual search.

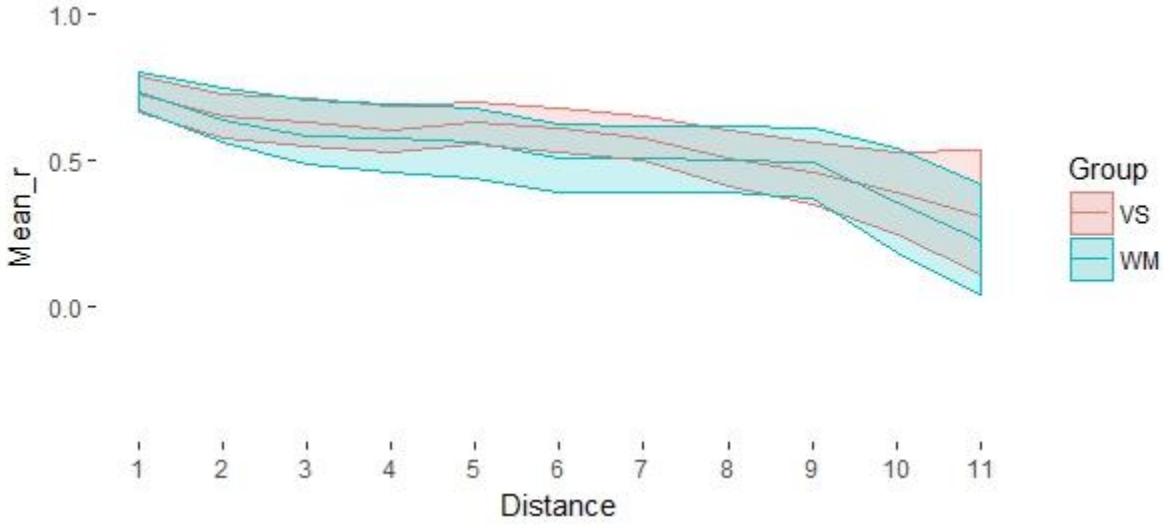

Figure 7. Quasi-simplex pattern - SH. Mean Pearson correlation between session scores (Mean_r) as a Function of the temporal distance between sessions. Bands indicate 95% credible

intervals estimated using Bayesian bootstrapping with 1,000 samples. VS = visual search; WM = procedural working memory.

5. General Discussion

In the present work, we reanalyzed results from two published training studies. Our goal was to test individual-differences predictions associated with a hypothesis (H1) for explaining the discrepancy between robust improvements in WM training tasks coupled with absent or nearly absent far transfer effects to Gf. According to H1, Gf and WM share cognitive processes, but this sharing of processes declines in the course of training because participants develop task-specific skills that reduce the WM demand of the training task. H1 predicts a quasi-simplex pattern combined with a declining pattern of Gf loadings of the per-session WM-training scores.

Our results indicate a robust quasi-simplex pattern, as predicted. In other words, the correlations between per-session scores declined with increasing temporal distance between the sessions, indicating a gradual change in the rank-ordering of individuals. Contrary to the predictions of H1, however, we observed stability or even an increase in the Gf loadings of the per-session scores of the training tasks. Such a trend could potentially reflect changes in the reliability of the per-session scores, but this was not the case here. We additionally found that with training progression, the four training tasks in DSVB became similar to one another in their individual differences. Below we consider two *post hoc* hypotheses (H2 and H3) that account for these results.

According to H2, the bulk of evidence for shared processes between WM and Gf is correlational rather than reflecting causation and, thus, is open to alternative explanations. One alternative explanation is that Gf and WM performance both reflect executive attention, but differ in the specific emphasis imposed by their specific respective tasks, with WM tasks taxing

primarily maintenance and Gf tasks mainly demanding disengagement (Shipstead et al., 2016). In that scenario, the stable or even increasing Gf loadings of WM performance over the course of training could indicate that WM task practice leads to less reliance of the trained tasks on maintenance and a shift to stronger reliance on executive attention. To explain the lack of transfer we observed in SH and DSVB, we would then need to assume that executive attention was not trained, or the transfer tasks employed at posttest are not sensitive enough to pick up the improvements.

Another alternative explanation is that the correlations reflect common factors having a biological basis that contribute to the development of these abilities rather than common cognitive processes (Garlick, 2002). Specifically, Garlick suggests that the ability to form new synapses will help in the development of many skills even when these skills do not have any cognitive processes in common. He actually suggested that this hypothesis reconciles the apparent contradiction between two lines of evidence. One comes from the neuropsychological literature, indicating that brain damages may cause highly circumscribed deficiencies. The other comes from research on intelligence, showing that all cognitive abilities are positively correlated.

To appreciate this point concerning lack of causal relation, take for example a similar correlation in another domain: Individual differences in the grip strength of the right arm are highly (in the .90s) correlated with those in the left arm (Hanten et al., 1999). Nonetheless, this does not lead to a clear prediction that training the right-hand grip would have a substantial influence on the strength of the left-hand grip. Instead, the intuitive prediction is that there would be hardly any transfer. Following this line of reasoning, improving performance in WM-tasks is not expected to (strongly) affect Gf. It is expected to influence performance in similar yet untrained WM tasks, however (Harrison et al., 2013). Notably, H2 contrasts with some classic

task analyses (e.g., Carpenter, Just, & Shell, 1990) indicating cognitive-process sharing between Gf and WM tasks. Nonetheless, H2 gains support from more recent empirical work. Specifically, Salthouse and Pink (2008) showed that the pattern of correlations between WM and Gf did not follow the trend expected based on the shared-process idea, namely that the WM-Gf relation was independent of WM task demands (i.e., the correlations were roughly equally high for low-load and high-load WM items).

Figure 8 presents a structural model that describes H3, another *post-hoc* hypothesis that integrates the findings. According to H3, success in the training task in a given session is determined by both g'/Gf and specific abilities (S1-S*k*). The contribution of these abilities gradually changes, as reflected in the fact that the specific ability in a given session is influenced by the specific ability in the preceding session but also by new abilities. Importantly, it is possible (perhaps even likely) that the new abilities that keep coming into play are not abilities that resulted from training, but instead, are pre-existing abilities demanded by the newly developed skill and which were not required beforehand[i].

What would S1-Sk may be? We provide some speculations below. The two first suggestions are related to the hypothesis that participants develop task specific skills that may be described as 'strategies'. Indeed, most participants in DSVB reported having used at least some strategy by the end of training. One suggestion regards a task-specific skill involving the retrieval of previous solutions from memory. Relying on this skill implies a potential contribution of individual differences in the rate and fidelity in which previously stored solutions can be used towards task performance. This ability may therefore be one of the specific abilities in Figure 8. Another ability involves whether one has identified the abstract common denominator in the four training tasks given to each group in DSVB. Thus, the ability to identify

a common denominator and make use of this knowledge could be another specific ability. Such ability may also explain the pattern of increased shared variance among the training tasks in SDVB. It might also explain the trend of increasing Gf loadings seen in that study, given the definition of intelligence as reflecting, among other things, the ability to understand complex ideas (Neisser et al., 1996), which in the present case is the abstract common denominator among the training tasks.

Our third speculation regarding the nature of S1-Sk is success motivation. This trait might influence individual differences at relatively late stages of training. Specifically, researchers who run relatively lengthy WM training studies are familiar with the fact that compliance with the training protocol demands commitment and motivation, and if serious steps are not taken, dropout rates may be high. In the present studies, dropout rates were quite small. Nonetheless, it is conceivable that some of these participants who stayed in the study despite a wish to leave it may have nonetheless lost their success motivation and performed poorly. Since success motivation may contribute to success on intelligence tests (Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011), it is conceivable that what differentiated between successful and less successful participants became gradually more motivation-related with training progression. In an initial attempt to examine this motivation hypothesis, we took advantage of the fact that, in DSVB, motivation was assessed by self-reports. However, the correlations between motivation and task success were generally very low, and more importantly, did *not* show the expected systematic gradual increase with training progression (see Supplemental Online Materials).

The fact that the Gf loadings remained stable (or even tended to increase) in the course of training does not contradict the hypothesis outlined in Figure 8, since Gf explained up to about

50% of the true variance in Session scores (Figure 1, DSVB). This figure of ~50% is based on the estimated loadings coupled with lack of evidence for meaningful reliability attenuation of these loadings. Specifically, one correlate, Gf was a latent variable and hence not contributing to reliability attenuation. The other correlate(s) were the Session scores that were nearly perfectly reliable. Given that the estimated loadings reached up to ~.70, this implies ~50% shared reliable variance. Most importantly, this further implies that there were about 50% of reliable variance are explained by specific abilities, such as task-specific skills.

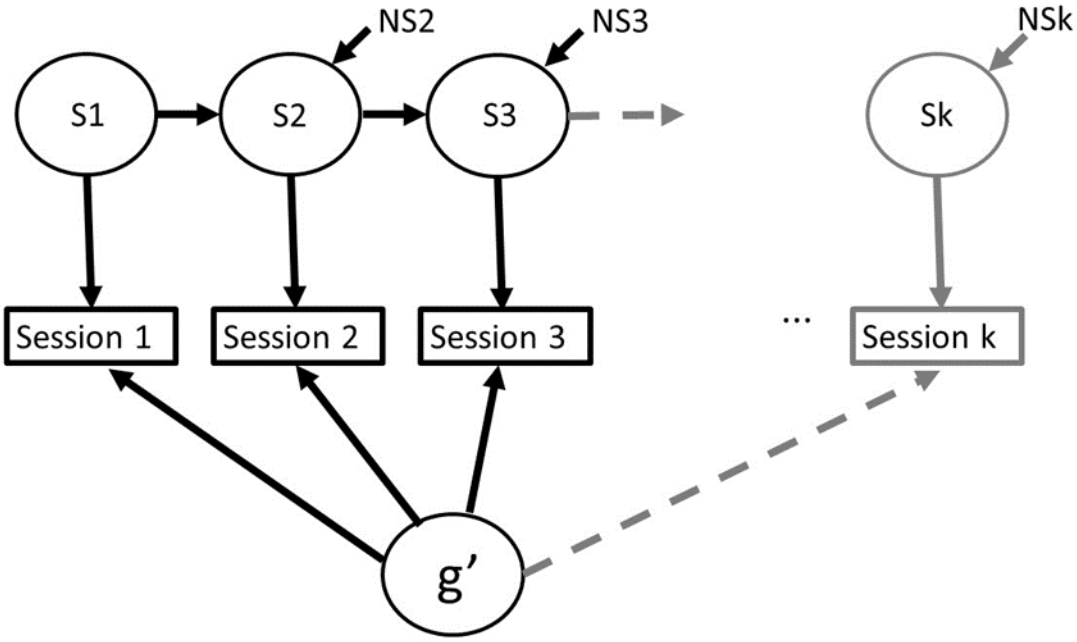

Figure 8: A post-hoc interpretation of the results. Session 1 till Session k represent success scores in the training task in the respective session. S1, S2.. Sk represent specific (non-Gf) abilities. NS=New Specific Ability. See text for details.

Before discussing some broader implications, we note several shortcomings of the present work. Perhaps the most serious shortcomings include the fact that Gf was poorly

operationalized in SH, and the relatively small *N*s as compared to usual applications of Structural Equation Modeling. Another obvious limitation concerns the fact that we studied two specific training methods, and whether the conclusions extend to other training protocols is something that should be tested in future studies.

In DSVB, we found a trend for increasing Gf loading in all the three groups. Furthermore, in that study, the results clearly supported the absence of training-related improvements in Gf. We therefore suggest that a better interpretation of the results is that with increasing training, the pre-existing individual differences in Gf became more visible, i.e., high Gf possibly contributed to the identification of a common abstract principle across four training tasks. This suggestion is in line with the positive relationship between initial ability and rate of improvement over training (Guye, De Simoni, & von Bastian, 2017). Such a trend supports in turn a recent criticism of "correlated gains" analyses that are sometimes used to support WM training efficacy (Tidwell, Dougherty, Chrabaszcz, Thomas, & Mendoza, 2014). Specifically, according to correlated-gains, success in far transfer is inferred from the fact that participants who showed large improvements in the training task also tended to show increased pre-to-post improvement. The criticism of correlated-gains analysis is precisely that individual differences in gains merely reveal pre-existing ability differences. Our results add another aspect to this criticism. Specifically, gain scores are based on subtraction (e.g., success in the last training session minus that in the first training session). Such subtraction may be justified if the session scores reflected the same makeup of abilities. However, the quasi-simplex shows this is not the case, making the difference difficult to interpret. A recent study (Sabah, Dolk, Meiran, & Dreisbach, 2018) indeed shows that examining training gains leads to paradoxical results,

whereby groups who have not showed improvement during training showed less near-transfer costs than groups who improved during training.

Perhaps the most important implication is that the present results point to the possibility that the attempt to improve Gf by improving WM rests on an assumption (regarding sharing of cognitive processes between Gf and WM) that is at best, unwarranted. We wish to emphasize the fact that we do not argue against the attempt to improve WM, which is worthwhile in its own right. We only suggest that the hope for such distant transferability that would reach Gf may prove unrealistic.

In conclusion, the present work addressed the question why WM training fails (or nearly fails) to transfer to Gf tasks. Re-analyses of results from two training studies indicate stable pattern or even an increasing pattern of Gf loadings with training progression. We additionally found that the correlations between session scores on the training task declined with increasing temporal distance between the sessions, reflecting a quasi-simplex pattern. These results contradict our *a-priori* H1: the hypothesis that WM training fails to generalize to Gf because performance on the training tasks gradually becomes less Gf-related. In fact, our findings point to the possibility that no such cognitive process-sharing exists in the first place (H2 and to some extent also H3). This conclusion suggests that the attempt to improve Gf through improving WM may be doomed to fail because it rests on unwarranted assumptions.

6. References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General*, *117*(3), 288–318. https://doi.org/10.1037/0096-3445.117.3.288

Alvares, K. M., & Hulin, C. L. (1972). Two explanations of temporal changes in ability-skill relationships: A literature review and theoretical analysis. *Human Factors*, *14*(4), 295–308. https://doi.org/10.1177/001872087201400402

Arthur, J., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices test. *Educational and Psychological Measurement*, *54*(2), 394–403. https://doi.org/10.1177/0013164494054002013

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2014). Improving fluid intelligence with training on working memory: a meta-analysis. *Psychonomic Bulletin & Review*, 1–12. https://doi.org/10.3758/s13423-014-0699-x

Bååth, R. (2016). bayesboot: An implementation of Rubin's (1981) Bayesian bootstrap (Version 0.2.1). Retrieved from https://cran.r-project.org/web/packages/bayesboot/index.html

Coghill, D. R., Seth, S., Pedroso, S., Usala, T., Currie, J., & Gagliano, A. (2014). Effects of methylphenidate on cognitive functions in children and adolescents with Attention-Deficit/Hyperactivity Disorder: Evidence from a systematic review and a meta-analysis. *Biological Psychiatry*, *76*(8), 603–615. https://doi.org/10.1016/j.biopsych.2013.10.005

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, *24*(1), 87–114; discussion 114-185.

Dahlin, E., Neely, A. S., Larsson, A., Bäckman, L., & Nyberg, L. (2008). Transfer of learning after updating training mediated by the striatum. *Science*, *320*(5882), 1510–1512. https://doi.org/10.1126/science.1155466

Dahlke, J. A., Kostal, J. W., Sackett, P. R., & Kuncel, N. R. (2018). Changing abilities vs. changing tasks: Examining validity degradation with test scores and college performance

criteria both assessed longitudinally. *Journal of Applied Psychology*, *103*(9), 980–1000. https://doi.org/10.1037/apl0000316

Dalal, R. S., Bhave, D. P., & Fiset, J. (2014). Within-person variability in job performance: A theoretical review and research agenda. *Journal of Management*, *40*(5), 1396–1436. https://doi.org/10.1177/0149206314532691

De Simoni, C., & von Bastian, C. C. (2018). Working memory updating and binding training: Bayesian evidence supporting the absence of transfer. *Journal of Experimental Psychology: General*, *147*(6), 829–858. https://doi.org/10.1037/xge0000453

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, *108*(19), 7716–7720. https://doi.org/10.1073/pnas.1018601108

Dunning, D. L., & Holmes, J. (2014). Does working memory training promote the use of strategies on untrained working memory tasks? *Memory & Cognition*, *42*(6), 854–862. https://doi.org/10.3758/s13421-014-0410-5

Dwyer, P. S. (1937). The determination of the factor loadings of a given test from the known factor loadings of other tests. *Psychometrika*, *2*(3), 173–178. https://doi.org/10.1007/BF02288394

Ekstrom, R. B., French, J. W., Harman, H. H., & Derman, D. (1976). *Kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Fischman, E. (1982). *Itellectual differential aptitude test battery*. Holon, Israel: Center for Technological Education.

Fleishman, E. A. (1972). On the relation between abilities, learning, and human performance. *American Psychologist*, *27*(11), 1017–1032. https://doi.org/10.1037/h0033881

Fleishman, E. A., & Rich, S. (1963). Role of kinesthetic and spatialvisual abilities in perceptual-motor learning. *Journal of Experimental Psychology*, *66*, 6–11.

Gallant, S. N. (2016). Mindfulness meditation practice and executive functioning: Breaking down the benefit. *Consciousness and Cognition*, *40*, 116–130. https://doi.org/10.1016/j.concog.2016.01.005

Guye, S., De Simoni, C., & von Bastian, C. C. (2017). Do individual differences predict change in cognitive training performance? A latent growth curve modeling approach. *Journal of Cognitive Enhancement*, *1*(4), 374–393. https://doi.org/10.1007/s41465-017-0049-9

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*, 0956797613492984. https://doi.org/10.1177/0956797613492984

Humphreys, L. G. (1960). Investigations of the simplex. *Psychometrika*, *25*(4), 313–323. https://doi.org/10.1007/BF02289750

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

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. https://doi.org/10.1073/pnas.0801268105

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. https://doi.org/10.1037/0278-7393.32.4.749

Keil, C. T., & Cortina, J. M. (2001). Degradation of validity over time: A test and extension of Ackerman's model. *Psychological Bulletin*, *127*(5), 673–697. https://doi.org/10.1037/0033-2909.127.5.673

Klingberg, T. (2010). Training and plasticity of working memory. *Trends in Cognitive Sciences*, *14*(7), 317–324. https://doi.org/10.1016/j.tics.2010.05.002

Kool, W., & Botvinick, M. (2014). A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology: General*, *143*(1), 131–141. https://doi.org/10.1037/a0031048

Laine, M., Fellman, D., Waris, O., & Nyman, T. J. (2018). The early effects of external and internal strategies on working memory updating training. *Scientific Reports*, *8*(1), 4045–4045. https://doi.org/10.1038/s41598-018-22396-5

Lewandowsky, S., Oberauer, K., Yang, L.-X., & Ecker, U. K. H. (2010). A working memory test battery for MATLAB. *Behavior Research Methods*, *42*(2), 571–585. https://doi.org/10.3758/BRM.42.2.571

Lövdén, M., Bäckman, L., Lindenberger, U., Schaefer, S., & Schmiedek, F. (2010). A theoretical framework for the study of adult cognitive plasticity. *Psychological Bulletin*, *136*(4), 659–676. https://doi.org/10.1037/a0020080

Meiran, N., Pereg, M., Kessler, Y., Cole, M. W., & Braver, T. S. (2015). The power of instructions: Proactive configuration of stimulus–response translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 768–786. https://doi.org/10.1037/xlm0000063

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer." *Perspectives on Psychological Science*, *11*(4), 512–534. https://doi.org/10.1177/1745691616635612

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian Structural Equation Models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1–30. https://doi.org/doi: 10.18637/jss.v085.i04

Minear, M., Brasher, F., Guerrero, C. B., Brasher, M., Moore, A., & Sukeena, J. (2016). A simultaneous examination of two forms of working memory training: Evidence for near transfer only. *Memory & Cognition*, 1–24. https://doi.org/10.3758/s13421-016-0616-9

Neath, A. A., & Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 199–203. https://doi.org/10.1002/wics.199

Neisser, U., Boodoo, G., Bouchard Jr., T. J., Wade, A., Brody, N., Ceci, S. J., … Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*(2), 77–101. https://doi.org/10.1037/0003-066X.51.2.77

Nevo, B., & Oren, C. (1986). Concurrent validity of the American Scholastic Aptitude Test (SAT) and the Israeli inter-University Psychometric Entrance Test (IUPET). *Educational and Psychological Measurement*, *46*(3), 723–725. https://doi.org/10.1177/0013164486463029

Oberauer, K. (2005). Binding and inhibition in working memory: individual and age differences in short-term recognition. *Journal of Experimental Psychology: General*, *134*(3), 368–387. https://doi.org/10.1037/0096-3445.134.3.368

Oberauer, K., & Hein, L. (2012). Attention to information in working memory. *Current Directions in Psychological Science*, *21*(3), 164–169. https://doi.org/10.1177/0963721412444727

R Core Team. (2014). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rabbitt, P. (2004). *Methodology of frontal and executive function*. Psychology Press.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., … Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology: General*, *142*(2), 359–379. https://doi.org/10.1037/a0029082

Revelle, W. (2017). psych: Procedures for psychological, psychometric, and personality research (Version 1.7.8). Retrieved from https://cran.r-project.org/web/packages/psych/index.html

Sabah, K., Dolk, T., Meiran, N., & Dreisbach, G. (2018). When less is more: Costs and benefits of varied vs. fixed content and structure in short-term task switching training. *Psychological Research*. https://doi.org/10.1007/s00426-018-1006-7

Sala, G., Tatlidil, K. S., & Gobet, F. (2018). Video game training does not enhance cognitive ability: A comprehensive meta-analytic investigation. *Psychological Bulletin*, *144*(2), 111–139. https://doi.org/10.1037/bul0000139

Schmiedek, F., Lovden, 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*. https://doi.org/10.3389/fnagi.2010.00027

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shahar, N., & Meiran, N. (2015). Learning to control actions: Transfer effects following a procedural cognitive control computerized training. *PLoS ONE*, *10*(3), e0119992. https://doi.org/10.1371/journal.pone.0119992

Shahar, N., Pereg, M., Teodorescu, A. R., Moeran, R., Karmon-Presser, A., & Meiran, N. (2018). Formation of abstract task representations: Exploring dosage and mechanisms of working memory training effects. *Cognition*.

Shahar, N., Teodorescu, A. R., Usher, M., Pereg, M., & Meiran, N. (2014). Selective influence of working memory load on exceptionally slow reaction times. *Journal of Experimental Psychology: General*, *143*(5), 1837–1860. https://doi.org/10.1037/a0037190

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working memory capacity and fluid intelligence: Maintenance and disengagement. *Perspectives on Psychological Science*, *11*(6), 771–799. https://doi.org/10.1177/1745691616650647

Szmalec, A., Verbruggen, F., Vandierendonck, A., & Kemps, E. (2011). Control of interference during working memory updating. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 137–151. https://doi.org/10.1037/a0020365

Tidwell, J. W., Dougherty, M. R., Chrabaszcz, J. R., Thomas, R. P., & Mendoza, J. L. (2014). What counts as evidence for working memory training? Problems with correlated gains and dichotomization. *Psychonomic Bulletin & Review*, *21*(3), 620–628. https://doi.org/10.3758/s13423-013-0560-7

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. https://doi.org/10.3758/BF03192720

von Bastian, C. C., & Eschen, A. (2016). Does working memory training have to be adaptive? *Psychological Research*, *80*(2), 181–194. https://doi.org/10.1007/s00426-015-0655-z

von Bastian, C. C., Langer, N., Jäncke, L., & Oberauer, K. (2013). Effects of working memory training in young and old adults. *Memory & Cognition*, *41*(4), 611–624. https://doi.org/10.3758/s13421-012-0280-7

von Bastian, C. C., Locher, A., & Ruflin, M. (2013). Tatool: A Java-based open-source programming framework for psychological studies. *Behavior Research Methods*, *45*(1), 108–115. https://doi.org/10.3758/s13428-012-0224-y

von Bastian, C. C., & Oberauer, K. (2013). Distinct transfer effects of training different facets of working memory capacity. *Journal of Memory and Language*, *69*(1), 36–58. https://doi.org/10.1016/j.jml.2013.02.002

von Bastian, C. C., & Oberauer, K. (2014). Effects and mechanisms of working memory training: a review. *Psychological Research*, *78*(6), 803–820. https://doi.org/10.1007/s00426-013-0524-6

Wilhelm, O., Hildebrandt, A. H., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00433

[i] The fit of the model presented in Figure 6 to the data could be estimated by using Structural Equations Modeling, for example. However, these analyses require much larger samples than we had, typically exceeding $N = 200$, which is why we had to settle for less direct analyses.