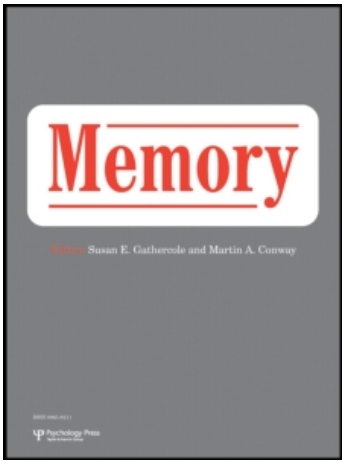

Memory
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713683358

The concurrent validity of the *N*-back task as a working memory measure

Susanne M. Jaeggi [a]; Martin Buschkuehl [a]; Walter J. Perrig [b];Beat Meier [b]
[a] University of Michigan, Ann Arbor, MI, USA [b] University of Bern, Berne, Switzerland

First published on: 19 April 2010

To cite this Article Jaeggi, Susanne M. , Buschkuehl, Martin , Perrig, Walter J. andMeier, Beat(2010) 'The concurrent validity of the *N*-back task as a working memory measure', Memory, 18: 4, 394 — 412, First published on: 19 April 2010 (iFirst)
To link to this Article: DOI: 10.1080/09658211003702171
URL: http://dx.doi.org/10.1080/09658211003702171

MEMORY, 2010, 18 (4), 394–412

Psychology Press
Taylor & Francis Group

The concurrent validity of the *N*-back task as a working memory measure

Susanne M. Jaeggi and Martin Buschkuehl
University of Michigan, Ann Arbor, MI, USA

Walter J. Perrig and Beat Meier
University of Bern, Berne, Switzerland

The *N*-back task is used extensively in literature as a working memory (WM) paradigm and it is increasingly used as a measure of individual differences. However, not much is known about the psychometric properties of this task and the current study aims to shed more light on this issue. We first review the current literature on the psychometric properties of the *N*-back task. With three experiments using task variants with different stimuli and load levels, we then investigate the nature of the *N*-back task by investigating its relationship to WM, and its role as an inter-individual difference measure. Consistent with previous literature, our data suggest that the *N*-back task is not a useful measure of individual differences in WM, partly because of its insufficient reliability. Nevertheless, the task seems to be useful for experimental research in WM and also well predicts inter-individual differences in other higher cognitive functions, such as fluid intelligence, especially when used at higher levels of load.

Keywords: Validity; Reliability; Inter-individual differences; Intelligence; Executive functions.

Working memory (WM) refers to the structures and processes used for temporarily storing and manipulating information in the face of ongoing processing and distraction. Although there are numerous ways to operationalise WM, one of the most popular measures of WM in neuroimaging literature is the *N*-back task (Conway et al., 2005; Kane & Engle, 2002). The reason to prefer the *N*-back task over traditional WM span tasks in functional neuroimaging lies in the appealing way to manipulate WM load and in its response requirements, which are less complex than in standard WM capacity tasks (Conway, Kane, & Engle, 2003). Typically, in the *N*-back task participants are presented with a stream of stimuli, and the task is to decide for each stimulus whether it matches the one presented *N* items before. It has been shown that the processing load can be varied systematically by manipulating the value of *N*, which is expressed with changes in accuracy and reaction time (RT) (see, e.g., Jonides et al., 1997). Despite its widespread use in neuroimaging, the psychometric properties of the *N*-back task as a WM measure have been rarely addressed. In addition, not much is known about individual differences in *N*-back performance and their relation to individual differences in other cognitive ability measures (cf. Jarrold & Towse, 2006). The goal of the present paper is to fill this gap as we aim to understand the nature of

Address correspondence to: Susanne M. Jaeggi, Department of Psychology, The University of Michigan, 530 Church Street, Ann Arbor, MI, 48109-1043, USA. E-mail: sjaeggi@umich.edu

The preparation of this manuscript was supported by a fellowship from the Swiss National Science Foundation (PA001-117473) to SMJ. We thank Daniela Denzler, Liliane Michlig, Axel Rau, Brigitte Schindler, Philipp Schmutz, and Lukas Schneider for their help with the data collection, as well as Priti Shah, Kristin Flegal, and Kirti Thummala for helpful comments on an earlier version of this paper.

http://www.psypress.com/memory DOI:10.1080/09658211003702171

the *N*-back task by investigating its relationship to WM and executive functions, and its role as an inter-individual difference measure.

The *N*-back task was originally introduced by Kirchner (1958) as a visuo-spatial task with four load factors (''0-back'' to ''3-back''), and by Mackworth (1959) as a visual letter task with up to six load factors. Gevins et al. (1990) introduced it to the field of neuroscience by using it as a ''visuomotor memory task'' with one load factor (3-back). The task involves multiple processes, such as the encoding of the incoming stimuli, the monitoring, maintenance, and updating of the material, as well as matching the current stimulus to the one that occurred *N* positions back in the sequence. Decision, selection, inhibition, and interference resolution processes are also involved (for a comprehensive task analysis, see Jonides et al., 1997, p. 471). The sequential nature of the task requires the execution of all those processes simultaneously, especially the simultaneous storage *and* processing of the material, which presumably led to the classification of the *N*-back task as a WM measure (Jonides et al., 1997; Kane & Engle, 2002). However, performance of the *N*-back task also seems to depend on processes that go beyond ''traditional'' WM-related processes. For example, it has been proposed that in the *N*-back task there are conflicting processes between familiarity and recollection; for example, if a current stimulus matches a previous stimulus, but not the one *N* items back in the sequence (Oberauer, 2005, p. 375). Resolving this conflict requires controlled processes, namely inhibition and interference resolution (Kane, Conway, Miura, & Colflesh, 2007). Further, Oberauer (2005) argued that binding processes are also involved, in that successful performance depends ''on the ability to establish and maintain bindings between the contents and their temporal context'' (p. 375).

Thus, the *N*-back task is a complex measure involving multiple processes that seem to be largely stimulus and material independent. In general, regardless of the material used, the number of errors as well as RTs increase monotonically with increasing levels of *N* (but not necessarily linearly; see Jaeggi, Schmid, Buschkuehl, & Perrig, 2009, for a discussion on this issue). On a neural level the results from neuroimaging studies are also consistent in revealing reliable activation increases in selected cortical areas with increasing processing load (e.g., Drobyshevsky, Baumann, & Schneider, 2006; Jonides et al., 1997; see Owen, McMillan, Laird, & Bullmore, 2005, for a meta-analysis). The areas most commonly showing this load-dependent activation change are primarily located in bilateral prefrontal and parietal cortices Those areas are part of a network that is commonly activated in WM tasks (e.g., Awh et al., 1996; see Wager & Smith, 2003, for a meta-analysis). Although these main areas of activation have been observed independent of the type of materials (Nystrom et al., 2000; Owen et al., 2005; Ragland et al., 2002; Schumacher et al., 1996), additional, stimulus-specific regions are also selectively activated (Knops, Nuerk, Fimm, Vohn, & Willmes, 2006; Owen et al., 2005).

Although behavioural and imaging results seem to be largely consistent, the psychometric properties of the *N*-back task remain largely unexplored. Only a few studies have addressed the psychometric properties, and some of them explicitly reported reliability measures of the *N*-back task in visual and verbal 0- to 3-back tasks (Friedman et al., 2006; Friedman et al., 2008; Hockey & Geffen, 2004; Kane et al., 2007; Oberauer, 2005; Salthouse, Atkinson, & Berish, 2003; Shamosh et al., 2008; Shelton, Elliott, Hill, Calamia, & Gouvier, 2009; Van Leeuwen, Ven den Berg, Hoekstra, & Boomsma, 2007). The reliability measures of these studies are mixed, ranging between $r = .02$ and $r = .91$, and in general, only higher task levels (2- and 3-back) seem to result in reliability estimates exceeding .80, presumably because of issues with ceiling performance in the lower levels. Given that the expected maximum correlation of a test with any other test is limited by its reliability, it is clear that the extent to which relationships with other variables can be established is restricted by the reliability of that measure itself (cf. Meier & Perrig, 2000).

Concerning construct validity, there are a few studies in which the *N*-back task has been correlated with other WM measures. What is notable here is that in those studies using a single WM capacity measure such as a reading span task (RST) or an operation span task report rather weak intercorrelations (ranging between $r = .10$ and $r = .24$; Colom, Abad, Quiroga, Shih, & Flores-Mendoza, 2008; Kane et al., 2007; Oberauer, 2005; Roberts & Gibson, 2002). Nevertheless, there are two studies (Shelton et al., 2009; Shelton, Metzger, & Elliott, 2007) who reported a correlation between operation span and *n*-back performance of $r \approx .46$ across three samples by using a composite *n*-back score consisting of

0-, 1-, 2-, and 3-back. Furthermore, by using a composite score of four complex span measures (operation span, reading span, symmetry span, and rotation span), Shamosh et al. (2008) obtained a correlation with a 3-back task of $r = .55$. Thus one could speculate that the low correlations in the former examples might result in part from too-large error variance by using just one performance measure. Interestingly however, it looks as though the *N*-back task is equally or even more closely related to simple span measures than to complex WM-span measures (correlations between $r = .12$ and $r = .53$; Colom et al., 2008; Dobbs & Rule, 1989; Gevins & Smith, 2000; Oberauer, 2005; Roberts & Gibson, 2002; Shelton et al., 2007, 2009). In addition, as more direct evidence of the relationship between various WM measures and *N*-back performance, we have shown that if people train on an *N*-back task they also improve performance in simple span tasks (digit span) after training, but not in a complex WM span measure (RST) (Jaeggi, Buschkuehl, Jonides, & Perrig, 2008). Thus the general pattern suggests that the *N*-back task is more closely related to simple than to complex WM span measures, a pattern that seems to stand in contrast to the face validity of the rather complex *N*-back task

However, there are other complex measures that seem to be related to *N*-back performance, such as classical measures of executive functions (EFs). In particular, EFs such as inhibitory control and set shifting seem to share a considerable amount of variance with *N*-back performance. Some authors have argued that the *N*-back task requires the retrieval of items that are no longer in the focus of attention, thus requiring a shift of attention (McElree, 2001; Verhaeghen & Basak, 2005; Verhaeghen, Cerella, & Basak, 2004). Indeed, a study with children (Ciesielski, Lesnik, Savoy, Grant, & Ahlfors, 2006) showed that 2-back performance is substantially correlated with Stroop performance ($r = .55$), Wisconsin Card Sorting ($r = -.56$), and verbal fluency ($r = .59$). However, other studies (Friedman et al., 2006, 2008) have reported only very weak correlations between 2-back and Stroop performance ($r = .10$, or $r = .12$, respectively).

Finally, several studies have looked at the relationship between the *N*-back task and intelligence. This is an important research question, since there is widespread evidence that WM shares considerable variance with measures of fluid intelligence (*Gf*) (Ackerman, Beier, & Boyle, 2005; Kane, Hambrick, & Conway, 2005; Kyllonen & Christal, 1990; Oberauer, Schulze, Wilhelm, & Suss, 2005). In addition, WM tasks and measures of *Gf* have been shown to recruit similar neural networks (Duncan et al., 2000; Gray, Chabris, & Braver, 2003; Kane & Engle, 2002). While there are many studies showing that especially complex WM span tasks predict inter-individual differences in measures of *Gf* (e.g., Engle, Tuholski, Laughlin, & Conway, 1999; Süss, Oberauer, Wittmann, Willhelm, & Schulze, 2002), there are also studies showing correlations between *N*-back performance and various intelligence measures (Friedman et al., 2006, 2008; Gevins & Smith, 2000; Salthouse, Pink, & Tucker-Drob, 2008; Shelton et al., 2009; Van Leeuwen et al., 2007; Waiter et al., 2009). Their correlation coefficients range between $r = .19$ and $r = .66$, suggesting shared variance between *N*-back performance and *Gf*. Further, with an inter-individual differences approach, Hockey and Geffen (2004) and Gevins and Smith (2000) investigated whether individual differences in intelligence (as measured with the Multidimensional Aptitude Battery or the Wechsler Adult Intelligence Scale, respectively) predict performance in the *N*-back task—and indeed, participants with high IQ scores were found to perform more quickly in the *N*-back task, especially at higher task levels. Finally, a more causal relationship between *N*-back performance and *Gf* was demonstrated by our training study, showing that people who train with the *N*-back task also improve performance in measures of *Gf* (Jaeggi et al., 2008).

Given this relationship between *N*-back performance and *Gf*, it seems counterintuitive that there is only a modest relationship between *N*-back and complex WM span measures, which in turn also predict inter-individual differences in *Gf*. The most parsimonious explanation for this might be that the RST and the *N*-back task each account for independent variance in *Gf* (Jaeggi et al., 2008; Kane et al., 2007). Nevertheless, it could also be that they share something in common that is not easily captured with inter-correlations, such as for example attentional control processes (Gray et al., 2003; Kane et al., 2004).

In this study we aim to shed more light on the nature of the *N*-back paradigm by means of three experiments investigating the relationship of the *N*-back task to WM and EFs, and also by investigating its role as an inter-individual

differences measure: In the first experiment we investigate the relationship between the *N*-back task and the RST. In Experiment 2 we look at the relationship between the *N*-back task and simple and complex WM tasks (digit span forward, digit span backward, and RST), and further explore its relation to a measure of EFs, an updating task (i.e., the self-ordered pointing task, SOPT; Petrides & Milner, 1982). In the last experiment we look at the validity of the *N*-back task as an inter-individual difference measure by investigating its relationship with *Gf*. In each experiment we varied stimulus materials and levels of load for the *N*-back task in order to disentangle their differential impact on the target measures. As an additional feature, we included dual-task versions of the *N*-back task because it has been proposed that dual tasks are usually purer estimates of WM capacity (WMC) because they prevent the use of strategies (Oberauer, Lange, & Engle, 2004), and further, because we have previously demonstrated that dual-task versions are well predictive of inter-individual differences in *Gf* (Jaeggi et al., 2008).

EXPERIMENT 1

Method

Participants. A total of 116 participants (55 women) took part in the experiment. They were recruited by undergraduate students in order to fulfil course credit, without any specified selection criteria apart from being native German speakers. Participants received no payment. The mean age was 29.09 years ($SD = 4.53$) and 95% of the participants had a college degree or higher level of education.

Apparatus. Task administration was computerised for the *N*-back task and run on a personal computer with a 17-inch display (resolution set to 1024×768 pixel) using the software package E-Prime (Psychology Software Tools, Pittsburgh, PA). Participants' responses were registered with a PST Serial Response Box (Psychology Software Tools, Pittsburgh, PA) with millisecond accuracy interfaced to the computer. The RST was administrated as paper and pencil test.

General procedure

All participants first answered sociodemographic questions regarding their age and their education-level. Participants then completed the *N*-back task and the RST, in counterbalanced order.

N-back task. The task and its material was the same as already used and described in previous studies (Jaeggi et al., 2007, 2008, 2009). We used visuospatial and auditory-verbal material as stimuli. The visuospatial stimuli consisted of blue squares appearing at one of eight different loci spaced equally and symmetrically around a constantly present white fixation cross in the centre of a black screen ($-7.16°/6.13°$; $0°/4.93$; $7.16/6.13°$; $-6.54°/0°$; $6.54°/0°$; $-7.16°/-6.13°$; $0°/-4.93°$; $7.16°/-6.13°$[1]; monitor-to-eyes distance 50 cm). The verbal material comprised eight aurally presented German consonants (c, g, h, k, p, q, t, w) spoken by a female voice set at a comfortable volume. Spatial locations and consonants were both chosen on the basis of their distinctiveness as assessed in pilot experiments.

The task was used with three levels of difficulty (1-back to 3-back) administered as single and dual tasks (see Figure 1). Participants were instructed to respond whenever the current stimulus was the same as the one presented *N* positions back in the sequence (*N* depending on the load level, that is, 1, 2, or 3).

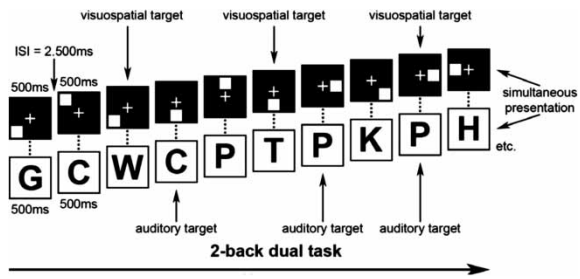

Figure 1. Example of the *N*-back task with the stimulus material used in Experiments 1 and 3. A response was required whenever the current stimulus matched the stimulus one, two, or three positions back in the sequence. The task was performed as a single task with auditory-verbal or visuospatial-nonverbal material only, but also as a dual task, where the attention had to be divided between two tasks presented simultaneously in each modality as shown in the example.

[1] A positive value indicates a location above or on the right side of the fixation cross and a negative value indicates a location below or on the left of the fixation cross.

Each trial consisted of a stimulus that was presented for 500 ms, followed by an interstimulus interval of 2500 ms, after which the next stimulus was presented. In the dual-task conditions the verbal and visuospatial stimuli were presented simultaneously, and participants had to independently process each modality, whereas the task difficulty (e.g., 2-back) was always the same for both modalities.

Participants performed three separate experimental blocks: a visuospatial, a verbal, and a dual-task block. Within each block there were three runs consisting of a single *n*-back load level, which lasted 2 minutes (40 trials) each. The stimuli were arranged in a pseudo-randomised order, i.e., the position of the targets was counterbalanced between the different runs. All runs were matched for the number of targets (33%) and non-targets (67%), as well as for distractors (e.g., 2-back targets in a 3-back run).

Participants were instructed to respond as quickly and accurately as possible. They were asked to press a specified button with their right index finger for targets in the auditory tasks, and another button with their left index finger for targets in the visuospatial tasks; no responses were required for non-targets. The same response allocation was used in the dual-task conditions, where the targets could occur in just one modality, in both modalities at the same time, or in none.

Each task condition was explained to the participant, followed by a few practice trials. Half of the participants started the experiment with the visuospatial block and then completed the auditory-verbal block. The other half performed the single tasks in the reverse order. All participants performed the dual-task block last. The runs within each block started with the 1-back task, followed by the 2-back and the 3-back task in that order. Performance was assessed in terms of reaction times (RTs; hits only) and accuracy (P_r, proportion hits minus false alarms; Snodgrass & Corwin, 1988) serving as dependent variables.

Reading span task (RST). The task consisted of 100 unrelated and relatively simple sentences (with 6 additional training sentences), which participants read aloud and indicated with ''yes'' or ''no'' for each sentence, whether it made sense semantically or not. Sentences were presented one by one on single paper sheets and were removed as soon as the participants had made their yes/no decision. Additionally, participants had to retain the last word of each sentence and recall these words in the correct order after presentation of two, three, four, five, or six sentences, whereas the amount of sentences corresponded to the level of difficulty. There were five sets per level. The material for the RST was provided by courtesy of Meredyth Daneman and translated into German by the first author. Each of the 100 sentences contained 6 to 15 words (*M*: 10.05; *SD*: 1.98) with a mean word length of 6.25 (*SD*: 0.81). Half of the sentences made sense semantically and half of them did not, but all of them were syntactically correct. We used a truncated method of administration (Friedman & Miyake, 2005)—i.e., the task was terminated after the participant did not reach a certain criterion of performance; that is, after the participant failed to recall any of the sets at a particular level. The dependent variable was defined as the highest level at which the participant recalled a majority of sets (three or more out of five). In addition, following the original scoring method of Daneman and Carpenter (1980), participants were given half a point for getting two out of five sets (e.g., if a participant recalled five sets at Level 2, four sets at Level 3, and two sets at Level 4, that participant would receive a span score of 3.5).

Results

Descriptive statistics. In the RST, participants reached an average score of 2.77 (*SD*: 0.84). For the N-back task, performance measures (means and standard deviations) and Spearman-Brown-corrected split-half reliability coefficients are presented in Table 1.

N-back task. We conducted $2 \times 2 \times 3$ repeated-measures analyses of variances (ANOVAs) with task condition (single vs dual-tasks), modality (visuospatial, auditory-verbal), and load (1-back to 3-back) as independent variables, and reaction time (RT; hits only) and accuracy (P_r; hits minus false alarms) as dependent variables.

As expected, there were significant main effects of load (1-back to 3-back), accuracy: $F(2, 230) = 1267.69$, $p < .001$, $\eta_p^2 = 0.92$; RTs: $F(1.57, 220) = 187.42$, $p < .001$, $\eta_p^2 = 0.63$; of task (single vs dual condition), accuracy: $F(1, 115) = 484.92$, $p < .001$, $\eta_p^2 = 0.81$; RTs: $F(1, 110) = 558.65$, $p < .001$, $\eta_p^2 = 0.83$; and of modality (visual-nonverbal vs auditory-verbal), accuracy: $F(1, 115) = 33.01$, $p < .001$, $\eta_p^2 = 0.22$; RTs: $F(1, 110) = 172.63$,

TABLE 1
Descriptive measures as well as split-half reliability for each variant of the *N*-back task of Experiments 1, 2 and 3.

	Experiment 1				*Experiment 2*				*Experiment 3*			
Accuracy (P_r) *Single tasks*	*M*	*SD*	*Range*	*r*	*M*	*SD*	*Range*	*r*	*M*	*SD*	*Range*	*r*
Visual												
1-back	0.99	0.03	0.78–1.00	0.93	0.98	0.06	0.50–1.00	0.75	0.96	0.09	0.48–1.00	0.18
2-back	0.91	0.13	0.48–1.00	0.85	0.91	0.10	0.40–1.00	0.45	0.92	0.09	0.71–1.00	0.26
3-back	0.66	0.19	0.08–1.00	0.51	0.75	0.15	0.30–1.00	0.45	0.60	0.22	0.11–1.00	0.51
Auditory												
1-back	0.98	0.04	0.85–1.00	−0.20	0.96	0.08	0.59–1.00	0.17	0.96	0.08	0.68–1.00	0.70
2-back	0.89	0.12	0.51–1.00	0.62	0.91	0.11	0.29–1.00	0.48	0.85	0.15	0.39–1.00	0.65
3-back	0.47	0.19	−0.08–1.00	0.39	0.67	0.16	0.20–0.98	0.44	0.45	0.21	0.00–0.92	0.47
Dual tasks												
1-back	0.94	0.06	0.74–1.00	0.11	0.83	0.14	0.11–1.00	0.58	0.85	0.17	0.05–1.00	0.74
2-back	0.72	0.17	0.27–1.00	0.63	0.63	0.15	0.17–0.91	0.55	0.62	0.18	0.23–0.90	0.59
3-back	0.33	0.14	0.04–1.00	0.41	0.40	0.15	0.06–0.76	0.60	0.29	0.17	−0.05–0.73	0.44
Reaction times (to hits only)												
Single tasks												
Visual												
1-back	507	176	245–1037	0.94	618	153	327–912	0.79	475	173	258–1042	0.96
2-back	547	183	253–1097	0.86	631	156	339–1034	0.71	540	197	287–1211	0.86
3-back	659	246	282–1398	0.69	774	209	357–1337	0.66	665	226	282–1194	0.68
Auditory												
1-back	592	140	300–1012	0.90	818	158	540–1458	0.54	595	176	355–1448	0.90
2-back	693	199	332–1289	0.69	904	191	517–1477	0.51	691	197	353–1348	0.66
3-back	1019	350	378–2019	0.54	1140	230	704–1759	0.28	955	356	417–1904	0.61
Dual tasks												
1-back	1052	280	561–1892	0.83	1173	222	644–1918	0.73	929	285	486–2108	0.83
2-back	1292	339	619–2052	0.74	1361	238	909–1983	0.57	1152	319	529–1801	0.50
3-back	1426	406	337–2643	0.45	1469	275	840–2095	0.48	1245	417	515–2157	0.66

Experiment 1: $N = 116$; Experiment 2: $N = 70$ for the single tasks, $N = 141$ for the dual task; Experiment 3: $N = 50$. *M*: mean; *SD*: standard deviation; *r*: split-half reliability (calculated as Pearson's correlation, corrected with the Spearman-Brown Prophecy Formula).

$p < .001$, $\eta_p^2 = 0.61$. All two-way interactions were significant: load × task, accuracy: $F(1.59, 230) = 68.81$, $p < .001$, $\eta_p^2 = 0.37$; RTs: $F(1.60, 220) = 17.04$, $p < .001$, $\eta_p^2 = 0.13$; load × modality, accuracy: $F(1.87, 230) = 31.45$, $p < .001$, $\eta_p^2 = 0.21$; RT: $F(1.45, 220) = 34.02$, $p < .001$, $\eta_p^2 = 0.24$; task × modality, accuracy: $F(1, 115) = 9.05$, $p < .01$, $\eta_p^2 = 0.07$; RT: $F(1, 110) = 11.46$, $p = .001$, $\eta_p^2 = 0.09$. There was also a significant three-way interaction (load × task × modality), accuracy: $F(1.50, 230) = 6.83$, $p < .01$, $\eta_p^2 = 0.06$; RTs: $F(1.40, 220) = 4.16$, $p < .05$, $\eta_p^2 = 0.04$). Post hoc tests showed that accuracy significantly dropped with increasing load level and that the single tasks were easier than the dual tasks (all $p < .01$; Bonferroni corrected for multiple comparisons). There was no difference between the auditory and the visuospatial tasks, except for the 3-back single-task versions, where performance was significantly better in the visuospatial version ($p < .01$); see Figure 2. For RTs, in addition to similar outcomes for the post-hoc tests as described for accuracy, there were additional modality differences on all load levels, i.e., participants reacted faster in response to the visuospatial than the auditory targets (all $p < .01$; Bonferroni corrected for multiple comparisons).

Correlation analysis. Pearson's product–moment correlations for all dependent measures are shown in Table 2. All in all, the various *N*-back conditions were moderately interrelated, which was more consistently expressed in RTs than in accuracy.

Regarding the correlations between the *N*-back task and the RST, which was of greater interest here, our results showed hardly any relation between the two tasks. Indeed, only the single auditory and visuospatial 3-back versions showed significant correlations with the RST (*auditory*: $r = .24$, *visuospatial*: $r = .19$), which were observed in RTs only. In terms of accuracy, the correlations between the *N*-back task and the RST were close to zero.

Discussion

First of all, in terms of performance, the load manipulation of our *N*-back task yielded very robust results in both modalities, with the auditory modality being even more sensitive than the visuospatial modality to load manipulations, as indicated by longer response latencies and more errors as the levels of *N* increased. Even greater demands are placed on the processing system if the task is conducted as dual task, and again this was more pronounced in the auditory modality regarding latencies. The intercorrelations between the various *N*-back conditions suggest that, despite their differences in load and modality, they seem to rely on related processing mechanisms.

Our variants of the *N*-back task confirm mixed reliability reports of earlier studies (e.g., Hockey & Geffen, 2004; Oberauer, 2005; Salthouse et al., 2003; Van Leeuwen et al., 2007). In general, the reliability measures were higher in respect to RTs than to accuracies, which is consistent with reports of Hockey and Geffen (2004). Concerning RTs, the highest reliabilities were observed in the single visuospatial 1-back task ($r = .94$). In the

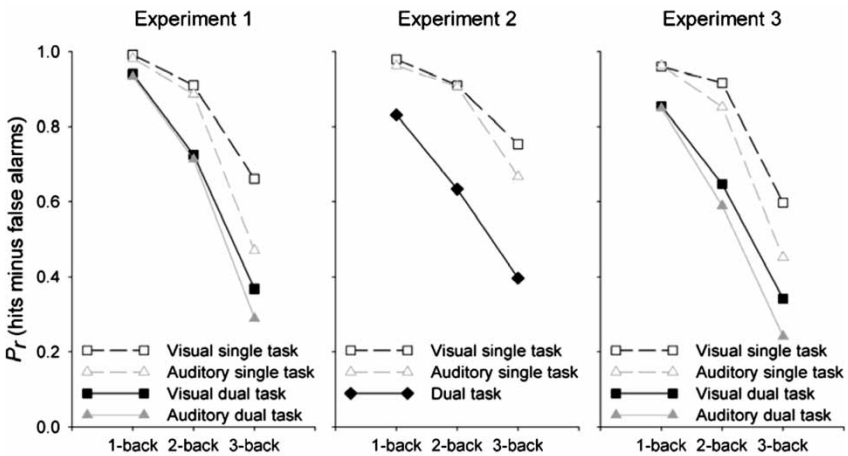

Figure 2. Performance (accuracy) for each *N*-back version for the three levels of load for Experiment 1 ($N = 116$), Experiment 2 ($N = 70$ for each of the single tasks, $N = 141$ for the dual task), and Experiment 3 ($N = 50$).

TABLE 2
Pearson's correlation coefficients for Experiment 1

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*
1 RST																			
N-back task (RTs)																			
Single tasks																			
Visuospatial																			
2 1-back	0.04																		
3 2-back	0.06	**0.65**																	
4 3-back	**0.19**	**0.47**	**0.54**																
Auditory																			
5 1-back	0.05	**0.75**	**0.55**	**0.40**															
6 2-back	0.02	**0.50**	**0.49**	**0.45**	**0.58**														
7 3-back	**0.24**	**0.20**	**0.28**	**0.36**	**0.25**	**0.39**													
Dual tasks																			
8 1-back	**0.01**	**0.30**	**0.32**	**0.28**	**0.44**	**0.50**	**0.31**												
9 2-back	−0.11	**0.19**	**0.38**	**0.39**	**0.35**	**0.39**	**0.41**	**0.76**											
10 3-back	−0.06	0.04	**0.27**	**0.33**	**0.24**	**0.32**	**0.29**	**0.64**	**0.69**										
N-back task (Acc)																			
Single tasks																			
Visuospatial																			
11 1-back	−0.05	−0.02	−0.12	−0.01	0.03	0.11	0.03	0.14	0.10	0.12									
12 2-back	−0.10	0.14	0.04	0.00	0.11	0.03	0.12	0.15	0.14	0.19	**0.37**								
13 3-back	−0.08	−0.16	**−0.24**	−0.11	−0.18	−0.09	−0.13	−0.09	−0.03	−0.06	0.13	**0.30**							
Auditory																			
14 1-back	−0.05	−0.04	−0.06	0.14	0.03	0.10	0.10	0.16	0.16	0.18	0.62	0.37	0.18						
15 2back	−0.14	−0.02	−0.02	−0.05	−0.11	**−0.42**	−0.18	−0.14	−0.10	0.02	0.00	0.11	0.15	0.11					
16 3-back	0.03	−0.01	−0.01	0.00	−0.09	**−0.19**	0.04	−0.10	−0.04	0.06	0.08	0.18	0.18	−0.02	**0.27**				
Dual tasks																			
17 1-back	0.00	−0.05	−0.07	0.01	−0.16	−0.09	0.10	**−0.29**	−0.15	−0.08	0.11	0.04	0.09	0.15	0.08	0.15			
18 2-back	−0.02	0.02	0.02	0.10	−0.01	0.00	**0.20**	−0.02	0.10	0.18	**0.19**	**0.33**	**0.38**	**0.32**	**0.34**	**0.41**	**0.30**		
19 3-back	0.05	0.08	0.07	0.12	0.08	−0.04	0.00	−0.04	0.05	0.10	0.13	0.15	**0.35**	0.16	**0.29**	**0.41**	**0.32**	**0.42**	

$N = 116$. Significant correlations are indicated in bold (*$p < .05$; 2-tailed).

most difficult 3-back conditions, however, the reliability was modest. Since only RTs of hits were entered in the analysis, the resulting low reliability estimates could be attributed to the fewer number of data points in this condition, which also applies to the dual task versions in general. Concerning accuracy, the 2-back tasks were the most reliable in general (with the exception of the highly reliable visuospatial 1-back task). The lower reliability estimates in the 3-back versions might again have resulted from increased error variance. In contrast, the low reliabilities in the easy 1-back tasks most likely reflect ceiling effects.

However, the main purpose of Experiment 1 was to explore the relationship between the *N*-back task and one of the standard measures of WMC, the RST (Conway et al., 2005; Daneman & Carpenter, 1980). Consistent with prior literature (e.g., Kane et al., 2007; Oberauer, 2005; Roberts & Gibson, 2002), we found no substantial correlation between the RST and any version or load level of the *N*-back task. The only exceptions were modest correlations with RTs in the single auditory-verbal and visuospatial versions of the 3-back task. Thus, by only looking at the correlations, the results suggest that performance of the *N*-back task and the RST result from different sources of variance, and thus they do not seem to load on the same WM construct at all (see also Jaeggi et al., 2008; Kane, 2005). However, the *N*-back task's low reliability might also have obscured a stronger relationship.

In the next experiment we thus further explored the question of WMC and its relation to the *N*-back task. The *N*-back task used in Experiment 1 only consisted of unrelated consonants as stimuli, and thus its verbal demands were presumably very low. This feature could have obscured the relationship with the RST, since this measure is sometimes specifically related to language-related abilities (MacDonald & Christiansen, 2002) (however, see Kane et al., 2004). Therefore, in Experiment 2 we used a modified version of the *N*-back in various load levels that required more language-related processes than the version used in Experiment 1. Again, we correlated the participants' performance on this task with the RST. Further, we used the digit span task (DST; forward and backwards version) as a simple span measure in order to investigate the differential relationship of the *N*-back task to simple and complex WM measures. Finally, we included the self-ordered pointing task (SOPT; Petrides & Milner, 1982) as an additional executive task involving controlled processes and memory updating. The SOPT was chosen because the *N*-back task is commonly seen as involving memory updating (e.g., Oberauer, 2005; Salthouse et al., 2003) and we therefore assumed a close relationship between those two tasks.

EXPERIMENT 2

Method

Participants. In this experiment, 281 participants (208 women) took part in exchange for course credit at the Department of Psychology at the University of Bern. The mean age was 21.89 years ($SD = 2.53$) and all were native German speakers.

Design. For this experiment we chose a between-participants design; i.e., we administered three versions of the *N*-back task, which were varied between groups: the first group ($N = 70$) performed only an auditory version, and the second group ($N = 70$) performed a visuospatial version. The last group ($N = 141$) performed a dual task version, combining the above-mentioned single tasks. All other tasks (RST, DST, and SOPT) were the same for all participants.

Apparatus. Task administration was computerised for all tasks apart from the DST and run on a Windows-based computer programmed with E-prime (Psychology Software Tools, Pittsburgh, PA). Participants' responses were registered with a standard computer mouse and a standard keyboard.

The following task description follows the order of task administration in the experiment.

N-back task. For this experiment we used a modified version of the *N*-back task, which included verbal material in both modalities with the addition of a spatial component: The auditory stimuli consisted of concrete one-syllable words spoken by either a male or a female voice, which were presented separately in either the left or the right ear via headphones. The visual stimuli consisted of easily nameable objects presented at one of four different locations on the computer screen. Participants responded to stimuli on the keyboard on specified keys: the left index finger was required in response to visual targets and the

right index finger to auditory targets. For the auditory stimuli a response was required whenever the current stimulus matched the side of presentation, the voice, *and* the spoken word with the stimulus presented *N* trials back in the sequence. For the visual stimuli a response was required whenever the stimuli matched the location *and* the object with the one presented *N* trials back in the sequence. No responses were required to non-targets. In the dual-task conditions stimuli were presented simultaneously and responses had to be made for each modality independently. The *N*-back load was always the same in both modalities. Similar to Experiment 1, the stimuli were presented for 500 ms with an ISI of 2500 ms. After some practice trials, all participants did a first run consisting of a 1-back task, a 2-back task, and a 3-back task, and a second run in the same order. There were 66 trials per *N*-back load with 22 targets presented in a pseudorandom sequence. RT (hits only) and accuracy (P_r; hits minus false alarms) served as dependent variables.

Self ordered pointing task (SOPT). The SOPT was originally developed by Petrides and Milner (1982) and is commonly regarded as neuropsychological measure of the capacity to initiate behaviour, and of the monitoring and organisation of this behaviour by means of strategies and plans. The task requires the self-initiated organisation and execution of a sequence of answers instead of a mere reproduction of sequences, and is thus classified as executive (Bryan & Luszcz, 2001). It also involves monitoring and updating, since participants have to continuously keep track of the answers that they have already given and to monitor the remaining possible answers. There has been some debate in the literature about whether the SOPT relies heavily on high-level WM processes (Daigneault & Braun, 1993; Petrides & Milner, 1982), as some authors found little evidence for the involvement of WM processes in the task (Bryan & Luszcz, 2001). Participants were presented with a display of 12 stimuli (4 horizontal, 3 vertical; 200×200 pixel) meant to be difficult to verbalise (i.e., faces; material as used by Lehmann et al., 2004), and which were randomly selected from a set of 36 stimuli for each participant. The participants were required to click on as many different stimuli as possible without clicking on the same stimulus twice, a mistake that was indicated by a beep. After each click the arrangement of the stimuli was changed randomly by the program. The task was terminated as soon as the participant made more than three mistakes. After some practice trials with concrete shapes, participants performed two runs with 12 stimuli each, the first with male faces only and the second with female faces only. The number of stimuli participants clicked on until the first error was made was taken as performance measure, and the mean of this score for both runs served as dependent variable.

Reading span task (RST). The same material, procedure, and scoring method were used as in Experiment 1, but this time the task was administered partly computerised: sentences were presented sequentially on a computer screen and controlled for presentation time. After two, three, four, five, or six sentences, participants were requested to recall the last words of each sentence in the correct order. The control of the presentation time was added, since there was evidence from Experiment 1 that performance in recall was related to the amount of time participants spent with reading (and presumably engaging in additional rehearsal). Thus sentence length was determined by median split according to which short sentences were presented for 8100 ms and long sentences for 8600 ms, followed by a blank screen that remained until the participant responded to indicate whether the sentence was semantically correct. The presentation time was determined by pilot experiments with a separate student sample ($N = 54$).

Digit-span task (DST). The DST was conducted in a forward and a backward condition following the procedure of the HAWIE-R (Tewes, 1991)—i.e., the German version of the WAIS (Wechsler, 1981). There were two trials per digit list length, and the dependent measure represents the total number of correctly reproduced trials of digits prior to failing two consecutive trials at a particular set size.

Results

Descriptive statistics: WM and updating tasks. Since all three participant groups performed the same WM and updating tasks, the following measures reflect the means and standard deviations for the whole group (N = 281). RST performance was comparable to Experiment 1 (*M*: 2.93; *SD*: 0.99), and digit span performance (forward digit span: *M*: 7.97; *SD*: 1.92; backwards digit

span: *M*: 7.60; *SD*: 2.02) was within normal range considering the education level (Tewes, 1991). For the SOPT, the performance score was *M*: 8.87 (*SD*: 1.53).

N-back tasks. Means and standard errors of performance scores and RTs for each task variant, as well as Spearman-Brown-corrected split-half reliability coefficients are presented in Table 1. Repeated-measures ANOVAs with *N*-back load as a within-participants factor were performed separately for each group of participants; i.e., for each *N*-back manipulation (auditory, visual, dual). As in Experiment 1, the results showed that performance significantly declined as the *N*-back level increased, which was evident in increasing RTs—*auditory*: $F(2, 138) = 104.03$; $p < .001$; $\eta_p^2 = 0.60$; *visual*: $F(2, 138) = 52.18$; $p < .001$; $\eta_p^2 = 0.43$; *dual*: $F(2, 248) = 120.59$; $p < .001$; $\eta_p^2 = 0.49$) and in decreasing accuracy (P_r; *auditory*: $F(2, 138) = 201.39$; $p < .001$; $\eta_p^2 = 0.75$; *visual*: $F(2, 138) = 125.78$; $p < .001$; $\eta_p^2 = 0.65$; *dual*: $F(2, 280) = 738.85$; $p < .001$; $\eta_p^2 = 0.84$. Performance measures indicated that the auditory version was slightly more difficult than the visual version, and the dual-task version was clearly more difficult than the single tasks (see Figure 2).

Correlation analysis. The intercorrelations in the *N*-back task are presented in Table 3. They can be ranked as mostly moderate to high, especially those concerning the RTs.

However, the pattern of correlations for the *N*-back task with the other WM and updating tasks is comparable with Experiment 1 (see Table 3): The RST hardly correlates with any of the *N*-back task versions, neither as single- nor as dual-task condition. On the other hand, the DST appears to be most closely correlated with the *N*-back task—especially in the visual single version, with values between $r = -.20$ and $r = .42$ (RTs), but also in all 3-back versions in accuracy, where *n*-back performance correlated between $r = .17$ and $r = .30$. The SOPT only modestly correlated with the *N*-back task (largest correlation: $r = 0.26$, 3-back auditory single task).

Finally, the intercorrelations between the RST, DST, and the SOPT were not large either, and in particular the SOPT did not substantially correlate with either WM measure, which is in line with findings of Bryan and Luszcz (2001).

Discussion

As in Experiment 1, the reliability estimates were mixed and generally higher for the RTs. Nevertheless, the load manipulation in the *N*-back task yielded very robust results both in terms of RT and accuracy. Error rates and response latencies increased with the level of *N*, which is most pronounced in the dual-task condition.

The results from the correlation analyses again show that the *N*-back task and other WM measures do not seem to share much common variance: Although the *N*-back material in this experiment was more explicitly verbal, and thus closer to the material used in the RST, the correlations between the *N*-back task and the RST were in the same (low) range as in Experiment 1. This pattern suggests that the general processes underlying *N*-back performance are independent of stimulus material, i.e., the relationship between the *N*-back task and the RST does not change by changing the material of the *N*-back task.

In contrast, and consistent with previous literature (e.g., Gevins & Smith, 2000; Roberts & Gibson, 2002), the simpler WM measure, the DST, seems to be more closely related to *N*-back performance than the complex WM task. The largest correlations between the DST and the *N*-back task were observed in response latencies in the visual version of the *N*-back task. The negative correlations indicate that faster RTs in the visual *N*-back tasks were related to better performance in the DST, which was more pronounced in the forward than the backward condition of the DST. The reason for this could lie in the more serial nature of the forward version that has more in common with the serial presentation of the *N*-back stimuli than the backward version of the DST.

In terms of accuracy, all significant correlations were obtained at highest level of load (3-back). Interestingly, the correlation coefficients for the *N*-back task did not differ to a large extent between the forward and the backward condition of the DST, although numerically it seems that the forward version is more closely related to the *N*-back task. It has been argued that the DST task taps mainly passive storage processes, which are relatively well practiced and automatised and thus require little executive processing (e.g., Engle et al., 1999; Gregoire & Van der Linden,

TABLE 3
Pearson's correlation coefficients for Experiment 2

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*
1 RST																						
2 DSF	**0.19**																					
3 DSB	**0.26**	**0.47**																				
4 SOPT (Acc)	0.09	0.09	−0.02																			
N-back task (RTs)																						
Single tasks																						
Visuospatial																						
5 1-back	−0.15	**−0.40**	−0.20	−0.08																		
6 2-back	−0.07	**−0.35**	−0.22	0.01	**0.78**																	
7 3-back	−0.17	**−0.42**	**−0.33**	0.02	**0.59**	**0.73**																
Auditory																						
8 1-back	0.04	0.06	0.06	−0.08																		
9 2-back	−0.01	−0.08	−0.02	−0.17				**0.68**														
10 3-back	−0.02	−0.06	−0.13	0.05				**0.38**	**0.55**													
Dual tasks																						
11 1-back	0.09	−0.04	−0.09	**0.18**																		
12 2-back	0.08	0.09	−0.06	**0.20**							**0.74**											
13 3-back	−0.11	−0.13	−0.17	0.00							**0.57**	**0.54**										
N-back task (Acc)																						
Single tasks																						
Visuospatial																						
14 1-back	−0.07	0.10	0.19	−0.06	−0.05	−0.07	0.03															
15 2-back	0.13	0.01	0.18	0.16	−0.13	−0.21	−0.03							**0.56**								
16 3-back	0.14	0.17	**0.27**	0.05	−0.11	−0.23	**−0.32**							**0.29**	**0.51**							
Auditory																						
17 1-back	0.03	0.08	0.06	0.02				**−0.66**	**−0.45**	**−0.29**							**0.38**					
18 2-back	0.03	0.10	0.00	0.19				**−0.38**	**−0.51**	−0.15							**0.33**	**0.60**				
19 3-back	0.17	**0.30**	**0.29**	**0.26**				**−0.24**	**−0.30**	−0.18												
Dual tasks																						
20 1-back	0.13	−0.08	0.02	0.10							**−0.30**	−0.13	−0.07							**0.38**		
21 2-back	0.14	0.03	0.10	−0.11							**−0.31**	**−0.20**	−0.01							**0.38**		
22 3-back	0.07	**0.25**	**0.21**	−0.12							**−0.35**	**−0.23**	**−0.25**							**0.33**	0.6	

RST: Reading span task; DSF: Digit span forward; DSB: Digit span backwards; SOPT: Self-ordered pointing task. RST, DSF, DSB, SOPT: $N = 281$; Auditory and visuospatial single *N*-back tasks: $N = 70$; dual *N*-back task: $N = 141$. Significant correlations are indicated in bold (*$p < .05$; 2-tailed).

1997; Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001)—hence they have been referred to as simple span measures. This might also be true for successful performance in the *N*-back task, which seems to be largely based on externally triggered recognition processes (Kane et al., 2007; Oberauer, 2005). On the other hand, the RST and the SOPT are usually considered high-level WM and executive tasks (Bryan & Luszcz, 2001; Just & Carpenter, 1992; Verhaeghen & Basak, 2005; Whitney, Arnett, Driver, & Budd, 2001). The executive processes mediating performance in the RST and the SOPT presumably reflect more self-generated and largely self-paced processes, and in addition the use of individual strategies (Bryan & Luszcz, 2000); processes that might be less crucial for successful performance in simple span and *N*-back tasks.

A further question is whether *N*-back performance also mediates performance in other higher-order cognitive tasks, such as for example *Gf* tasks, especially since we found improvements in *Gf* after training on *N*-back (Jaeggi et al., 2008). Specifically, we were interested in the role of the *N*-back task as an inter-individual differences measure. Thus in Experiment 3 we correlated the same variants of the *N*-back task as used in Experiment 1 with a measure of *Gf*, Raven's Advanced Progressive Matrices (RAPM; Raven, 1990). The RAPM is widely acknowledged as good measure of *Gf* (Conway et al., 2005; Cowan et al., 2005) and it also loads highly on a general factor (Spearman's *g*) in psychometric studies of intelligence (Carroll, 1993). In accordance with earlier studies (Gray et al., 2003; Kane et al., 2007), we expected substantial correlations between the two measures, and further, we predicted that the correlation between the *N*-back task and the intelligence measure should increase with increasing load (Hockey & Geffen, 2004; Salthouse et al., 2008; Stankov & Crawford, 1993). As dual tasks seem to be better predictors of *Gf* (Fogarty & Stankov, 1982; Spilsbury, 1992), we expected larger correlations of *Gf* and *N*-back performance in the dual task versions, also because of the larger inter-individual variability. Finally, since *Gf* is assumed to reflect a domain-free process (e.g., Kane & Engle, 2002), the correlations between the RAPM and the visuospatial and the auditory versions of the *N*-back task should be comparable.

EXPERIMENT 3

Method

Participants. A total of 50 participants (26 women; mean age: 20.44; *SD*: 3.56) volunteered to take part in the experiment. The participants were recruited by undergraduate students in order to fulfil course credit and were mainly undergraduate students. Participants received no payment.

N-back task. The same *N*-back task with the same material was used as described in Experiment 1.

Raven's APM. The RAPM (Raven, 1990) was developed as a measure of *Gf* in order to test participants with above-average intellectual abilities. In the version we used there are two sets of problems: Set I and Set II. We used Set I as a practice set, and only the data from Set II are included in the analyses. Set II consists of 36 visual analogy problems arranged by increasing difficulty. Each problem consists of a 3×3 matrix of patterns in which one pattern is missing. The participants are required to select the missing pattern from among eight response alternatives. Participants were provided with 40 minutes to work on the problems, and the dependent variable consisted of the number of correct solutions produced in that time.

Results

Descriptive statistics. Participants reached the expected range of performance in the APM (*M*: 27.9: *SD*: 4.03). Performance measures (means and standard deviations) as well as Spearman-Brown-corrected split-half reliability coefficients for the N-back tasks are presented in Table 1.

N-back task. As in Experiment 1, we conducted $2 \times 2 \times 3$ repeated-measures within-participants ANOVAs with task condition (single vs dual-tasks), modality (visuospatial, auditory-verbal), and load (1-back to 3-back) for RT and accuracy.

The overall results generally replicate the findings of Experiment 1; i.e., there were significant main effects of load (1-back to 3-back), accuracy: $F(1.48,\ 98) = 288.05$, $p < .001$, $\eta_p^2 = 0.85$; RTs: $F(1.58,\ 94) = 70.44$, $p < .001$, $\eta_p^2 = 0.60$); of task (single vs dual condition), accuracy: $F(1,\ 49) = 138.59$, $p < .001$, $\eta_p^2 = 0.74$; RTs: $F(1,\ 47) = 195.10$,

$p < .001$, $\eta_p^2 = 0.80$; and of modality (visual-nonverbal vs auditory-verbal), accuracy: $F(1, 49) = 24.28$, $p < .001$, $\eta_p^2 = 0.33$; RTs: $F(1, 47) = 71.50$, $p < .001$, $\eta_p^2 = 0.60$. The following two-way interactions were significant: load ×task, accuracy: $F(2, 98) = 16.89$, $p < .001$, $\eta_p^2 = 0.26$; RTs: $F(1.66, 94) = 5.48$, $p < .01$, $\eta_p^2 = 0.10$; load × modality, accuracy: $F(1.69, 98) = 10.85$, $p < .001$, $\eta_p^2 = 0.18$; RT: $F(1.64, 94) = 11.27$, $p < 001$, $\eta_p^2 = 0.19$; task × modality, accuracy: *ns*, $\eta_p^2 = 0.01$; RT: $F(1, 47) = 11.85$, $p = .001$, $\eta_p^2 = 0.20$. The three-way interaction (load × task × modality) was not significant, neither for accuracy nor for RT.

Post hoc tests again indicated that accuracy significantly dropped with increasing load level, that the single tasks were easier than the dual tasks (all $p < .01$; Bonferroni corrected for multiple comparisons), and also that there was no difference between the auditory and the visuospatial tasks, except for the 3-back single-task versions, where performance was significantly better in the visuospatial version ($p < .01$); see Figure 2. For RTs there were additional modality differences on all load levels, which were, however, only significant in the single-task conditions; i.e., participants reacted faster in response to the visuospatial than the auditory targets (all $p < .01$; Bonferroni corrected for multiple comparisons).

Correlation analysis. The full correlation matrix with the RAPM and all variants of the *N*-back task is presented in Table 4. First of all, the intercorrelations of the various *N*-back task conditions are comparable to Experiment 1. They are larger in regard to RTs than accuracy, and there are no correlations between accuracy and RTs in the single tasks.

There are no significant correlations between the *N*-back task and the RAPM in regard to RTs. However, there are significant correlations with accuracy measures on these tasks, with the largest correlation in the 3-back dual task version ($r = .48$).

Discussion

As in the preceding experiments, the load manipulation in the *N*-back task was very robust, with the auditory modality suffering more from the load manipulations, especially at the highest level of difficulty. Again, the dual-task conditions were clearly the most difficult variants. The results of intercorrelations between the various *N*-back conditions are largely consistent with Experiment 1. Consistent with Experiment 1 and 2, the reliability estimates are mixed overall, but again larger for RTs, especially in the single-task conditions, where the values are almost identical with those obtained in Experiment 1.

The correlation of the *N*-back task with *Gf* as assessed with the RAPM shows that there is at least some shared variance between the two tasks, which replicates findings of Gray et al. (2003) and Kane et al. (2007) as well as our previous study (Jaeggi et al., 2008). Our prediction that the correlation between the *N*-back task and the intelligence measure should be most pronounced at higher load levels (Hockey & Geffen, 2004; Stankov & Crawford, 1993) proved to be correct, suggesting that *Gf* and *N*-back performance are primarily related through attentional control processes (Gray et al., 2003; Kane et al., 2004), which are more pronounced at higher level of load (e.g., Smith & Jonides, 1997). However, the prediction that the highest correlations should be observed in the dual-task conditions (Fogarty & Stankov, 1982; Spilsbury, 1992) was only partly supported: Whereas the highest correlation was indeed observed in the 3-back dual-task condition, no correlations were observed in the 2-back dual-task versions, a pattern for which we do not have an explanation.

In sum, the results of Experiment 3 provide further evidence that performance in the *N*-back is related to inter-individual differences in *Gf*. Together with the results of Experiments 1 and 2, our data suggest that the *N*-back task and WM spans contribute differentially to performance in *Gf* tests, a conclusion that is consistent with Kane et al. (2007).

GENERAL DISCUSSION

The aim of the present experiment was to investigate the psychometric properties of the *N*-back task, to shed light on its relation with other WM and EF tasks, and to investigate its role as an individual difference measure. Although the *N*-back task has been used almost exclusively in the context of WM, it's role as a "pure" WM measure is being discussed (Cowan, 2001; Kane et al., 2007; Oberauer, 2005). The three experiments presented here provide additional evidence in this direction by showing that the *N*-back task

TABLE 4
Pearson's correlation coefficients for Experiment 3

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*
1 Fluid Intelligence(Raven)																			
N-back task (RTs)																			
Single tasks																			
Visuospatial																			
2 1-back	−0.16																		
3 2-back	−0.10	**0.84**																	
4 3-back	−0.19	**0.62**	**0.68**																
Auditory																			
5 1-back	−0.23	**0.81**	**0.84**	**0.61**															
6 2-back	−0.15	**0.64**	**0.78**	0.66	0.80														
7 3-back	−0.10	**0.51**	**0.55**	**0.58**	**0.59**	**0.54**													
Dual tasks																			
8 1-back	0.01	**0.63**	**0.61**	**0.44**	**0.57**	**0.55**	**0.40**												
9 2-back	0.02	**0.61**	**0.61**	**0.56**	**0.53**	**0.59**	**0.46**	**0.75**											
10 3-back	−0.07	**0.46**	**0.58**	**0.54**	**0.47**	**0.52**	**0.48**	**0.71**	**0.71**										
N-back task (Acc)																			
Single tasks																			
Visuospatial																			
11 1-back	−0.02	0.00	−0.05	−0.18	0.08	−0.10	0.20	0.02	−0.10	0.02									
12 2-back	0.26	0.04	−0.05	0.14	−0.08	−0.16	0.15	0.05	0.04	−0.02	0.12								
13 3-back	**0.32**	−0.03	−0.12	−0.08	−0.03	0.00	0.07	−0.06	−0.11	−0.09	0.26	**0.40**							
Auditory																			
14 1-back	−0.13	−0.04	−0.05	0.02	−0.01	−0.03	0.02	−0.03	−0.06	0.02	**0.34**	0.26	0.15						
15 2-back	0.17	0.05	0.05	0.08	0.06	−0.06	0.04	−0.11	−0.16	−0.05	**0.30**	**0.35**	0.25	0.21					
16 3-back	**0.29**	0.03	−0.01	−0.14	0.00	−0.11	−0.06	0.07	−0.07	−0.17	**0.28**	0.01	0.25	−0.08	**0.36**				
Dual tasks																			
17 1-back	−0.21	**−0.32**	**−0.38**	−0.17	**−0.29**	**−0.32**	**−0.29**	**−0.64**	**−0.36**	**−0.31**	0.08	0.07	0.06	0.24	0.11	−0.08			
18 2-back	0.04	**−0.30**	**−0.30**	−0.18	−0.24	**−0.30**	−0.13	**−0.36**	**−0.44**	−0.17	0.12	**0.29**	0.24	0.23	0.18	−0.01	**0.59**		
19 3-back	**0.48**	0.08	0.11	0.03	0.12	0.14	0.13	0.03	0.04	−0.05	0.12	0.10	**0.48**	**−0.28**	0.17	**0.45**	−0.02	0.16	

$N = 50$. Significant correlations are indicated in bold (*$p < .05$; 2-tailed).

is only weakly related to tasks that are commonly seen as complex WM capacity measures, such as the RST. The most parsimonious explanation for this finding would be that WM and/or executive functions are not unitary (Miyake et al., 2000; Salthouse et al., 2003; Stuss et al., 2002), and that the RST and the *N*-back task are not loading on the same WM or executive sub-component. However, the lack of correlation might also reflect the fact that the main processes that drive performance in the *N*-back tasks used here are familiarity- and recognition-based discrimination processes (Oberauer, 2005; Smith & Jonides, 1998). In contrast, in complex WM span tasks, rather than recognition, the main process seems to be active recall, thus the inter-relationship between the two tasks might be low because different processes are required (Kane et al., 2007). Indeed, Oberauer et al. (2005) reported correlations between a verbal *N*-back task and various recognition tasks that are considerably higher than the average correlations between complex span tasks. Furthermore, Shelton and colleagues (2007, 2009) used a modified recall version of the *N*-back task in their work. They obtained larger correlations with the operation span than did most other studies, thus, the overlap between *N*-back and complex WM span measure seems to be larger if both require similar recall processes (Kane et al., 2007, p. 621).

On the other hand, the mixed results regarding reliability in our studies—but also in previous research—make it difficult to draw any firm conclusions about the task's concurrent validity.

To conclude, the investigation of the concurrent validity of the *N*-back task with other cognitive measures suggests that the *N*-back task is a complex measure and that the processes involved are not easily disentangled. Looking at the data relating the *N*-back task to other measures of WMC, the *N*-back task does not seem to be a useful measure of individual differences in WMC, due to its low reliability. Nevertheless, it is a very useful tool for the experimental investigation of WM processes because it allows load to be manipulated in a very simple, straightforward way. Further, there is converging evidence that *N*-back performance can well predict individual differences in *Gf* and other higher cognitive functions, at least when used at higher levels of load.

Manuscript received 7 September 2009
Manuscript accepted 12 February 2010
First published online 19 April 2010

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.

Awh, E., Jonides, J., Smith, E. E., Schumacher, E. H., Koeppe, R. A., & Katz, S. (1996). Dissociation of storage and rehearsal in verbal working memory: Evidence from positron emission tomography. *Psychological Science*, *7*(1), 25–31.

Bryan, J., & Luszcz, M. A. (2000). Measurement of executive function: Considerations for detecting adult age differences. *Journal of Clinical and Experimental Neuropsychology*, *22*(1), 40–55.

Bryan, J., & Luszcz, M. A. (2001). Adult age differences in self-ordered pointing task performance: Contributions from working memory, executive function and speed of information processing. *Journal of Clinical and Experimental Neuropsychology*, *23*(5), 608–619.

Carroll, J. B. (1993). *Human cognitive abilities. A survey of factor-analytic studies*. Cambridge, MA: Cambridge University Press.

Ciesielski, K. T., Lesnik, P. G., Savoy, R. L., Grant, E. P., & Ahlfors, S. P. (2006). Developmental neural networks in children performing a categorical *N*-back task. *Neuroimage*, *33*(3), 980–990.

Colom, R., Abad, F. J., Quiroga, M. Á., Shih, P. C., & Flores-Mendoza, C. (2008). Working memory and intelligence are highly related constructs, but why? *Intelligence*, *36*, 584–606.

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.

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

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., et al. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, *51*, 42–100.

Daigneault, S., & Braun, C. M. (1993). Working memory and the Self-Ordered Pointing Task: Further evidence of early prefrontal decline in normal aging. *Journal of Clinical and Experimental Neuropsychology*, *15*(6), 881–895.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

Dobbs, A. R., & Rule, B. G. (1989). Adult age differences in working memory. *Psychology and Aging*, *4*(4), 500–503.

Drobyshevsky, A., Baumann, S. B., & Schneider, W. (2006). A rapid fMRI task battery for mapping of

visual, motor, cognitive, and emotional function. *Neuroimage*, *31*(2), 732–744.

Duncan, J., Seitz, R. J., Kolodny, J., Bor, D., Herzog, H., Ahmed, A., et al. (2000). A neural basis for general intelligence. *Science*, *289*(5478), 457–460.

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.

Fogarty, G., & Stankov, L. (1982). Competing tasks as an index of intelligence. *Personality and Individual Differences*, *3*, 407–422.

Friedman, N. P., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods*, *37*(4), 581–590.

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.

Friedman, N. P., Miyake, A., Young, S. E., Defries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General*, *137*(2), 201–225.

Gevins, A., & Smith, M. E. (2000). Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style. *Cerebral Cortex*, *10*, 829–839.

Gevins, A. S., Bressler, S. L., Cutillo, B. A., Illes, J., Miller, J. C., Stern, J., et al. (1990). Effects of prolonged mental work on functional brain topography. *Electroencephalography and Clinical Neurophysiology*, *76*(4), 339–350.

Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, *6*(3), 316–322.

Gregoire, J., & Van der Linden, M. (1997). Effects of age on forward and backward digit spans. *Aging, Neuropsychology, and Cognition*, *4*, 140–149.

Hockey, A., & Geffen, G. (2004). The concurrent validity and test–retest reliability of a visuospatial working memory task. *Intelligence*, *32*, 591–605.

Jaeggi, S. M., Buschkuehl, M., Etienne, A., Ozdoba, C., Perrig, W. J., & Nirkko, A. C. (2007). On how high performers keep cool brains in situations of cognitive overload. *Cognitive, Affective, and Behavioral Neuroscience*, *7*(2), 75–89.

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 of the United States of America*, *105*(19), 6829–6833.

Jaeggi, S. M., Schmid, C., Buschkuehl, M., & Perrig, W. J. (2009). Differential age effects in load-dependent memory processing. *Aging, Neuropsychology, and Cognition*, *16*, 80–102.

Jarrold, C., & Towse, J. N. (2006). Individual differences in working memory. *Neuroscience*, *139*(1), 39–50.

Jonides, J., Schumacher, E. H., Smith, E. E., Lauber, E. J., Awh, E., Minoshima, S., et al. (1997). Verbal working memory load affects regional brain activation as measured by PET. *Journal of Cognitive Neuroscience*, *9*(4), 462–475.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension. *Psychological Review*, *99*, 122–149.

Kane, M. J. (2005). Full frontal fluidity? Looking in on the neuroimaging of reasoning and intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 141–163). Thousand Oaks, CA: Sage.

Kane, M. J., Conway, A. R. A., 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), 615–622.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*(4), 637–671.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackermann, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), 66–71.

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.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352–358.

Knops, A., Nuerk, H. C., Fimm, B., Vohn, R., & Willmes, K. (2006). A special role for numbers in working memory? An fMRI study. *Neuroimage*, *29*(1), 1–14.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*, 389–433.

Lehmann, C., Mueller, T., Federspiel, A., Hubl, D., Schroth, G., Huber, O., et al. (2004). Dissociation between overt and unconscious face processing in fusiform face area. *Neuroimage*, *21*(1), 75–83.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review, 109*(1), 35–54; discussion 55–74.

Mackworth, J. F. (1959). Paced memorizing in a continuous task. *Journal of Experimental Psychology*, *58*(3), 206–211.

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 817–835.

Meier, B., & Perrig, W. J. (2000). Low reliability of perceptual priming: Consequences for the interpretation of functional dissociations between explicit and implicit memory. *Quarterly Journal of Experimental Psychology. A: Human Experimental Psychology*, *53*(1), 211–233.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex ''frontal lobe'' tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*(4), 621–640.

Nystrom, L. E., Braver, T. S., Sabb, F. W., Delgado, M. R., Noll, D. C., & Cohen, J. D. (2000). Working memory for letters, shapes, and locations: fMRI evidence against stimulus-based regional organization in human prefrontal cortex. *Neuroimage*, *11*(5 Pt 1), 424–446.

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.

Oberauer, K., Lange, E., & Engle, R. W. (2004). Working memory capacity and resistance to interference. *Journal of Memory and Language*, *51*(1), 80–96.

Oberauer, K., Schulze, R., Wilhelm, O., & Suss, H. M. (2005). Working memory and intelligence – their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), 61–65.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). *N*-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, *25*(1), 46–59.

Petrides, M., & Milner, B. (1982). Deficits on subject-ordered tasks after frontal and temporal-lobe lesions in man. *Neuropsychologia*, *20*, 249–262.

Ragland, J. D., Turetsky, B. I., Gur, R. C., Gunning-Dixon, F., Turner, T., Schroeder, L., et al. (2002). Working memory for complex figures: An fMRI comparison of letter and fractal n-back tasks. *Neuropsychology*, *16*(3), 370–379.

Raven, J. C. (1990). *Advanced Progressive Matrices. Sets I, II*. Oxford, UK: Oxford University Press.

Roberts, R., & Gibson, E. (2002). Individual differences in sentence memory. *Journal of Psycholinguistic Research*, *31*(6), 573–598.

Salthouse, T. A., Atkinson, T. M., & Berish, D. E. (2003). Executive functioning as a potential mediator of age-related cognitive decline in normal adults. *Journal of Experimental Psychology: General*, *132*(4), 566–594.

Salthouse, T. A., Pink, J. E., & Tucker-Drob, E. M. (2008). Contextual analysis of fluid intelligence. *Intelligence*, *36*, 464–486.

Schumacher, E. H., Lauber, E., Awh, E., Jonides, J., Smith, E. E., & Koeppe, R. A. (1996). PET evidence for an amodal verbal working memory system. *Neuroimage*, *3*(2), 79–88.

Shamosh, N. A., Deyoung, C. G., Green, A. E., Reis, D. L., Johnson, M. R., Conway, A. R., et al. (2008). Individual differences in delay discounting: Relation to intelligence, working memory, and anterior prefrontal cortex. *Psychological Science*, *19*(9), 904–911.

Shelton, J. T., Elliott, E. M., Hill, B. D., Calamia, M. R., & Gouvier, W. D. (2009). A comparison of laboratory and clinical working memory tests and their prediction of fluid intelligence. *Intelligence*, *37*, 283–293.

Shelton, J. T., Metzger, R. L., & Elliott, E. M. (2007). A group-administered lag task as a measure of working memory. *Behavior Research Methods*, *39*(3), 482–493.

Smith, E. E., & Jonides, J. (1997). Working memory: A view from neuroimaging. *Cognitive Psychology*, *33*(1), 5–42.

Smith, E. E., & Jonides, J. (1998). Neuroimaging analyses of human working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(20), 12061–12068.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50.

Spilsbury, G. (1992). Complexity as a reflection of the dimensionality of a task. *Intelligence*, *16*, 31–45.

Stankov, L., & Crawford, J. D. (1993). Ingredients of complexity in fluid intelligence. *Learning and Individual Differences*, *5*, 73–111.

Stuss, D. T., Alexander, M. P., Floden, D., Binns, M. A., Levine, B., McIntosh, A. R., et al. (2002). Fractionation and localization of distinct frontal lobe processes: Evidence from focal lesions in humans. In D. T. Stuss & R. T. Knight (Eds.), *Principles of frontal lobe function* (pp. 392–407). New York: Oxford University Press.

Süss, H-M., Oberauer, K., Wittmann, W. W., Willhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability – and a little bit more. *Intelligence*, *30*(3), 261–288.

Tewes, U. (1991). *Hamburg-Wechsler-Intelligenztest für Erwachsene-Revision 1991 (HAWIE-R)*. Bern: Hans Huber.

Van Leeuwen, M., Van den Berg, S. M., Hoekstra, R. A., & Boomsma, D. I. (2007). Endophenotypes for intelligence in children and adolescents. *Intelligence*, *35*, 369–380.

Verhaeghen, P., & Basak, C. (2005). Ageing and switching of the focus of attention in working memory: Results from a modified *N*-back task. *Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *58*(1), 134–154.

Verhaeghen, P., Cerella, J., & Basak, C. (2004). A working memory workout: How to expand the focus of serial attention from one to four items in 10 hours or less. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1322–1337.

Wager, T. D., & Smith, E. E. (2003). Neuroimaging studies of working memory: A meta-analysis.

Cognitive, Affective, and Behavioral Neuroscience, *3*(4), 255–274.

Waiter, G. D., Deary, I. J., Staff, R. T., Murray, A. D., Fox, H. C., Starr, J. M., et al. (2009). Exploring possible neural mechanisms of intelligence differences using processing speed and working memory tasks: An fMRI study. *Intelligence*, *37*, 199–206.

Wechsler, D. (1981). *WAIS-R Manual*. New York: The Psychological Corporation.

Whitney, P., Arnett, P. A., Driver, A., & Budd, D. (2001). Measuring central executive functioning: What's in a reading span? *Brain and Cognition*, *45*(1), 1–14.