Original Papers

Polish Psychological Bulletin
2016, vol. 47(1) 51–61
DOI - 10.1515/ppb-2016-0006

Konrad Kulikowski[*]
Katarzyna Potasz-Kulikowska[**]

Can we measure working memory via the Internet? The reliability and factorial validity of an online n-back task

Abstract*: The aim of this study was to check whether an online n-back task conducted in the uncontrolled environment of the Internet can yield valid and reliable data. For this purpose, 169 participants completed an online n-back task with n1, n2 and n3 blocks on their home computers. The results have shown acceptable reliability for overall accuracy and reaction time indices across n1, n2, n3 blocks, as well as for reaction time indices for each n block. Unacceptable reliability has been found for separate n levels accuracy indices and for response bias indices. Confirmatory factor analysis has revealed that, among 8 proposed measurement models, the best fit for the data collected is a model with two uncorrelated factors: accuracy consisting of n1, n2, n3 indices and reaction time consisting of n2, n3 indices. The results of this study have demonstrated for the first time that a reliable administration of online n-back task is possible and may therefore give rise to new opportunities for working memory research.*

Key words*: n-back, working memory, online, reliability, validity*

Psychologists can observe new or rare phenomena online and can do research on traditional psychological topics more efficiently, enabling them to expand the scale and scope of their research.

Report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet (Kraut et al., 2004, p. 105).

According to Eurostat (2015), in 2014 among European Union countries (EU 28), 81% of households had Internet access at home, and 75% of individuals (aged 16 to 74) used the Internet regularly, at least once a week. As the Internet becomes more popular, new opportunities for experimental psychology arise and in the future psychologists may move from local and stationary to global and online laboratories. The American Psychological Association generally approves Internet experimental research as "*inherently no more risky than traditional observational survey, or experimental methods*" (Kraut et al., 2004, p. 105).

Detailed discussion of issues and opportunities that arise from psychological research studies on the Internet is beyond the scope of this article but can be found elsewhere, e.g. in Gosling and Mason (2015). Generally there are empirical data suggesting that experimental research conducted via the Internet on participants' home computers might be reliable and valid. Probably the best known online psychological experiment is an Implicit Project run by Nosek, Banaji, and Greenwald (2002) to study implicit social cognition. Other noteworthy research studies are: an Alcohol Implicit Association Test demonstrated by Houben and Wiers (2008), an online continuous performance test introduced by Raz, Bar-Haim, Sadeh and Dan (2012), and an online computerized visual-spatial complex span task – the lion game – created by Van de Weijer-Bergsma, Kroesbergen, Prast and Van Luit (2014). Barchard and

[*] **Institute of Psychology, Jagiellonian University**
[**] **Department of Neurology, Jagiellonian University Medical College**

Corresponding author's contact information: Konrad Kulikowski, Institute of Psychology, Ingardena 6, 30-060 Kraków, konrad.kulikowski@uj.edu.pl
The authors declare that there are no conflicts of interest.

Williams (2008) stated that if the researchers are aware of the risks associated with Internet research, experiments in online environments can be conducted successfully.

Working memory

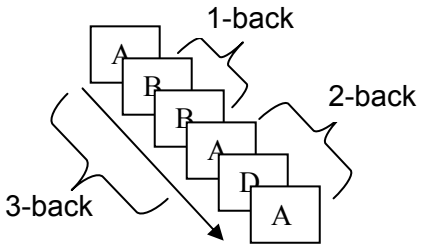
Working memory (WM) (Baddeley, 2012), a complex memory system responsible for processing, updating, maintaining and storing information (Oberauer, Süß, Wilhelm, & Wittman, 2003), is one of the most popular research topics among cognitive psychologists and neuroscientists. In the EBSCO database, more than 19,500 articles with *working memory* phrase in their titles can be found, and the classic article "Working Memory" by Baddeley and Hitch (1974) has more than 2,900 citations in Scopus. WM is considered an important correlate of a wide range of cognitive processes; studies show that WM is a predictor of: academic achievement (Colom, Escorial, Shih & Privado, 2007), multitasking performance (König, Bühner, & Mürling, 2005), language comprehension (Daneman & Merikle, 1996), the level of proficiency in a second language (Linck, Osthus, Koeth, & Bunting, 2014), effects of task-irrelevant sound on cognition (Sörqvist, 2010), reasoning ability (Ackerman, Beier, & Boyle, 2005; Oberauer, Schulze, Wilhelm, & Süss, 2005), probabilistic judgment (Dougherty & Hunter, 2003), creativity (De Dreu, Nijstad, Baas, Wolsink, & Roskes, 2012), explicative and predictive sentence comprehension (Pérez, Paolieri, Macizo, & Bajo, 2014), and controlled search of long-term memory (Unsworth, Brewer, & Spillers, 2012).

Experimental procedures that assess WM functioning are mainly designed as interactive computer scripts, which make them a natural candidate for online environments. The transfer of WM research to the Internet can lead to new research opportunities, resulting from large samples, diverse populations, and new big data exploration methods. However, the lack of control over the course of online experiments and technical issues (such as differences in computer CPU speed, operating systems, RAM memory, monitor size and refresh rates, speed of Internet access, and so on) raise the question of the reliability and validity of any data obtained. The main aim of this study is to empirically test the reliability and factorial validity of one of the most popular WM measurement paradigms – the n-back task, conducted online, outside of a lab, on participants' home computers. In comparison to other WM measurement methods, the n-back task is based on a relatively simply rule, has short trials, and is easy to perform on a computer. It has been used here as a well-established and popular example of a WM task, in order to answer the question of whether we can reliably conduct research on WM via the Internet on participants home computers. It is important to note that the main purpose of this study is to contribute to a debate on WM research conducted outside of labs, via the Internet, rather than to add to the debate on the reliability and validity of the standard n-back task.

The n-back task

N-back is a widely used WM measurement paradigm. In the n-back task (Kirchner, 1958; Cohen et al., 1997) a series of rapidly changing stimuli are presented one after another. Participants have to report whether or not the stimulus currently presented on the screen is the same stimulus that was presented n stimuli back (see Fig. 1). Participants have to recall the stimulus relevant to the current n level, prevent interference from other inadequate stimuli, and constantly update the relevant stimulus. The experimenter can manipulate the type of stimulus (letters, digits, shapes, etc.) as well as n numbers (2-back, 3-back, etc.) to observe how these changes may influence the storage and updating of information in the WM system.

Figure 1. Graphical illustration of 1, 2 and 3 n conditions in n-back paradigm with letter as a stimulus



The n-back task has four basic dependent measures: *false alarms* – number of reactions to non-target letters, *hits* – number of reactions to target letters, *misses* – number of omitted target letters, *correct rejections* – number of correct rejections of non-target letters. Due to the fact that the sum of hits and misses rates, as well as the sum of false alarms and correct rejections rates is 1, the false alarm rates and hits rates are sufficient to describe how participants react to both target and non-target stimuli.

In this study, to calculate all dependent measures, we have applied Two-High Threshold Model and correction recommended by Snodgrass and Corwin (1988), that is, we have added 0.5 to each frequency, and 1, to numbers of old or new trials. Thus we calculate the false alarms rate as: false alarms rate = (false alarms + 0.5)/(number of distracters + 1.0) and hits rate as: hits rate = (hits + 0.5)/(number of targets + 1.0). Then, following the Two-High Threshold Model (Snodgrass & Corwin, 1988), we calculate the accuracy index (Ac) as: accuracy = hits rate – false alarms rate. We subtract the false alarms rate from hits rate because false alarms are generated only for an uncertain state, thus they are a direct estimate of probability of hits when uncertain. So the hit rate (H) is composed of "true" accuracy (Ac) and false alarms (FA) as follows: H = Ac + FA; therefore, accuracy can be described by the formula: Ac = H – FA (Snodgrass & Corwin, 1988, p. 38). Additionally, as a dependent measure, we consider the reaction time (RT) for hits (RT has been used as an n-back measure in the work of Ragland et al., 2002, Jaeggi et al., 2010b, Hockey & Geffen, 2004) and response

bias calculated as: response bias = false alarms rate/(1.0 – accuracy). Ragland et al. (2002) suggest that accuracy can be considered a measure of performance success, and reaction time a measure of performance so we treat them as a two main dependent measures of interests.

The reliability and validity of the classic n-back task

To compare the reliability and validity of the online and standard versions of n-back, let us begin with a brief overview of research studies on the reliability and validity of the standard n-back task.

An analysis of how the same group of participants perform in an n-back task and another WM task, called complex span task, (Conway et al., 2005; Foster et al., 2014) is a method often used to test the validity of n-back (Redick & Lindsey, 2013); however, this approach gives inconclusive results. Kane, Conway, Miura and Colflesh (2007) suggest that "*N-back has face validity as a WM task, but it does not demonstrate convergent validity with at least 1 established WM measure*", and conclude that n-back results and complex span task results do not reflect the same WM construct. Miller, Price, Okun, Montijo, and Bowers (2009) also report that n-back is not a pure measure of WM, but can be used to assess general cognitive functioning of Parkinson's disease patients. Redick and Lindsey (2013), based on a meta-analysis, estimate the correlation between n-back and complex span task as r=0,20, and conclude that these two paradigms cannot be used interchangeably as WM measures. Shelton, Elliott, Hill, Calamia and Gouvier (2009) point out that n-back score may indicate an ability to update information stored in WM rather than reflect clear WM capacity.

In contrast to this, Schmiedek and colleagues used a latent variable approach, and found a statistically significant correlation: r=0,96 (Schmiedek, Hildebrandt, Lövdén, Lindenberger, & Wilhelm, 2009) and r=0,69 (Schmiedek, Lövdén, & Lindenberger, 2014) between a complex span factor and an updating factor represented by n-back task. Wilhelm, Hildebrandt, and Oberauer (2013) notice some advantages of n-back over complex span tasks, such as simplicity of scoring and short trial times, and argue that n-back tasks reflect the updating aspect of the general WM capacity construct and provide useful information for WM research.

N-back validity is still a matter of debate, and there is no common agreement on what is actually measured using the n-back paradigm.

Another controversy concerns n-back reliability. As presented in table 1, authors conducting n-back have used different n-back modes, different dependent measures and even different reliability estimate methods; thus, an analysis of n-back reliability gives mixed results. The n-back reliability coefficient, shown in table 1, ranges from 0,49 to 0,95; thus it is difficult to propose one general conclusion as to whether it is a reliable measure or not – it appears as if it may be reliable in some circumstances. Hockey and Geffen (2004) assessed n-back percentage accuracy scores as moderately reliable, and reaction times as highly reliable, but Jaeggi et al. (2010a) found that n-back tasks are insufficiently reliable. Jaeggi et al. (2010) concluded that n-back is not a useful measure of individual differences in WM, but can be useful for experimental research on WM, and as a predictor of interindividual functioning in higher cognitive functions.

Regardless of the n-back validity and reliability debate, the n-back paradigm is widely used in WM research. It has recently been used in studies on pain (Attridge, Noonan, Eccleston, & Keogh, 2015), childhood trauma (Philip et al., 2015), child neuropsychological development (Forns et al., 2014), neuropsychological assessment (Domínguez, Martín-Rodríguez, & León-Carrión, 2015), cognitive workload and fatigue (Guastello et al., 2015), chronic fatigue syndrome (Medow et al., 2014), the influence of antipsychotic drugs on WM functions (Goozee et al., 2015), intelligence training programs (Dougherty, Hamovitz & Tidwell, 2015), age-related decline in executive functions (Salminen, Frensch, Strobach, & Schubert, 2015), and in functional brain imaging research (Jacola et al., 2014).

The n-back task is not free from controversy, but it is a popular WM measurement paradigm used not only to measure WM functioning itself, but also to investigate the relationship between WM and other human characteristics. The introduction of an online n-back task may contribute to the further development of WM research, and possibly, in the future, to the debate over the reliability and validity of the n-back paradigm. In our opinion, testing online n-back reliability and factorial validity, might be important before introducing such a procedure in a research practice. Results of n-back task conducted in precisely controlled laboratory settings might be totally different when carried out on participants' home computers in uncontrolled online environments and without the presence of experimenter.

Method

Participants

One hundred and seventy-five psychology students from Jagiellonian University in Kraków took part in an online n-back task as part of the requirements of a cognitive psychology course. The responses of 6 participants were excluded due to missing data (listwise deletion was used); the results of 169 participants (30 male) were included, the mean age being 21 years (SD 2 years).

Procedure

We used a modified version of a single n-back task based on Jaeggi et al. (2010a), but instead of colored shapes we used a set of letters (20 consonants: B, C, D, F, G, H, J, K, L, M, N, P, Q, R, S, T, V, W, X, Z), letters were presented on the centre of the screen and have 20% of screen height. As in the study of Jaeggi et al. (2010a), we used n-back without lure trials, one of the n-back versions accepted in the literature (Redick, & Lindsey,

Table 1. Comparison of n-back paradigm reliability estimates

Authors	N-back mode	Dependent measures	Reliability estimates
Salthouse, Atkinson, & Berish (2003)	Verbally digits 1 back	Number of errors in repeating the sequence of digits	0,62sh
	Verbally digits 2 back	Number of errors in repeating the sequence of digits	0,77sh
Hockey & Geffen (2004)	Spatial 1 back	Accuracy / RT	0,49/0,79tr
	Spatial 2 back	Accuracy / RT	0,47/0,69tr
	Spatial 3 back	Accuracy / RT	0,73/0,80tr
Friedman et al. (2006)	Spatial 2 back	Proportion correct	0,91ca
Kane et al. (2007)	Letters 3 back	Proportion correct	0,84ca
Van Leeuwen, van den Berg, Hoekstra, & Boomsma (2007)	Spatial 3 back children	Number of correct responses	0,50tr
	Spatial 3 back adolescents	Number of correct responses	0,68tr
Friedman et al. (2008)	Spatial 2 back	Accuracy	0,91ca
Schmiedek et al. (2009)	Spatial 3 back	Accuracy	0,95ca
Jaeggi et al. (2010a)	Figural 2–4 back	Mean accuracy for 2+3 +4 back conditions	0,79ca
Jaeggi, Buschkuehl, Perrig, & Meier (2010b)	Spatial 1 back	Accuracy / RT	0,95/0,94sh
	Spatial 2 back	Accuracy / RT	0,85/0,86sh
	Spatial 3 back	Accuracy / RT	0,51/0,69sh
Schmiedek et al. (2014)	Numerical 3 back	Accuracy	0,92ca
	Spatial 3 back	Accuracy	0,95ca

ca – Cronbach's alpha, sh – split half correlation, tr – test retest, RT – reaction times for hits only, Accuracy – proportion of hits minus false alarms

2013). Lure foils occur when the stimulus does not match the n level that is currently presented, e.g. in a sequence B, L, B, M in n3 variant the second B will be lure, because it matches n2, not the n3 variant (Szmalec, Verbruggen, Vandierendonck, & Kemps, 2011). During the task white letters were shown on a black background for 500ms each, followed by a 2000ms interstimulus interval. Participants were asked to press the "A" key each time the current letter was identical to the one presented n positions back in a sequence, otherwise they were not to react. The response window lasted from the onset of a letter to the presentation of the next letter i.e. 2500ms. We used 1, 2, and 3 n-back levels in that order, and each n level was presented 3 times, so there were 9 blocks in total. There were 20 trials in each n1 block, 21 trials in each n2 block, and 22 trials in each n3 block. Each block consisted of 6 targets + 13 non-targets + n start trials. Start trials could not present the target letter, and their number depended on n, e.g. in a 2-back condition, we need at least 2 start trials, before the first target in a 2-back position can appear. Nine main blocks were preceded by self-paced instruction pages and practice blocks. During the practice, participants received one block consisting of 9 trials (6 non-targets, 3 targets) per each n level, so that there were 3 blocks in total. When practice blocks had been done, participants got the option to repeat the practice, and could repeat it as long as they thought they needed to. Participants received performance feedback in practice trials only. Inquisit 4 Web software (http://www.millisecond.com/about/publications.aspx) was used to create procedures and conduct the trials via the Internet. The n-back procedure was placed on Inquisit servers and participants received an email link to the procedure website.

When participants started the procedure, high-performance Inquisit engine (De Clercq, Crombez, Buysse, & Roeyers 2003) was downloaded locally to their machine in order to provide precise millisecond timing over the web.

At first, participants were informed about a study by teachers in a classroom. After this, they were asked by standardized e-mail message to complete the procedure on their home computers within a maximum of 7 days. According to our knowledge, this study is a first attempt to analyze reliability and factorial validity of n-back task performed on participants' home computers via the Internet. Thus, in order to minimize all potential unknown confounding variables influencing reliability other than online character of an experimental procedure, we decided to conduct this study on highly homogenous students' sample. The sample of students, previously asked in a classroom to participate in a study, has also the advantage of providing a high response rate. In this study we have sent 185 e-mails and we have received 175 responses, thus response rate is approximately 95%. According to Sánchez-Fernández, Muñoz-Leiva, and Montoro-Ríos (2012) there are two important stages of online research: (1) obtaining high response rate and (2) obtaining quality response. High response rate gives us a confidence that reliability of a task is influenced only by properties of experimental procedure, not a recruitment process or the composition of the sample. An obvious disadvantage of asking students in a classroom to participate in a study is the fact that this study was not fully online, because of a physical contact with a recruiter before the study. So, in fact, this study addresses the question: is the online n-back task reliable and valid procedure, when we provide the high response rate? Finally, it is worth noting, that online studies, similar to presented one, starting from information meeting and followed by conducting an online procedures on participants' home computers, are generally possible in the field of online research. We believe that using homogenous students' sample and providing high response rate is justified and allows us to minimize influence of confounding variables. This approach might also enable us to obtain more precise information about online n-back procedure reliability than a fully online study in which we simply put links to a procedure on the random websites.

Statistical analysis

In this study we intended to examine the reliability of WM functioning indicators which could be obtained from n-back task executed online on participants` home computers. Hence, we calculated and analyzed: hits rate, false alarms rate, accuracy, and response bias, mean and median reaction times to hits. Each of them was calculated in 5 variants: separately for results from n1, n2, n3 block, for the overall results from n2+n3 blocks and for the overall results from n1+n2+n3 blocks. Reliability was calculated as split half Pearson`s correlation corrected with the Spearman Brown formula, as a reliable we considered indicators which reliability coefficient was higher than 0,7. Factorial validity of measurement model was tested by confirmatory factor analysis (CFA) (Byrne, 2009) conducted in SPSS AMOS 18 supplemented by Composite Reliability (CR) (Raykov, 1997) and Average Variance Extracted (AVE) (Fornell & Larcker, 1981) indicators.

Results

Reliability

Detailed descriptive statistics and reliability estimates for hits rate, false alarms rate, accuracy, response bias, average reaction time and median reaction time are presented in table 2. The statistics are presented in five variants: separately for n1, n2, n3 block, for overall results from n2+n3 blocks, and for overall results from n1+n2+n3 blocks.

Hits

For hits rate (hits = correct reaction to a target stimulus) the reliability of indices decreases as the load on WM increases. N1 hits variant is the most reliable (0,86) from all 5 variants; however, in this case, we probably have to deal with a ceiling effect. For n1 variant, we can see a strong left-skewed distribution (skewness = -4,33) with high data concentration around the mean value (kurtosis = 23,15, SD = 0,08), this suggests that most participants have a high score in n1 variant, about 0,94 points. Due to the small variability in the n1 hits variant, the utility of this indicator may be marginal. N2 and n3 variants have reliability coefficients below 0,7, so that we cannot consider them to be reliable indices in contrast to aggregated indices n2+n3 and n1+n2+n3 which have an acceptable reliability with coefficients higher than 0,74.

False alarms

For false alarms rate (false alarms = reaction to a stimulus which is not the target), n1 has the lowest reliability (0,47) across 5 calculated variants; this may be due to low false alarms rate in n1 variant (mean = 0,03). In all likelihood, n1 false alarms variant scores reflect accidental, random mistakes made by participants, rather than as result of any cognitive mechanism. The reliability of n2 and n3 variants are low: 0,62 and 0,68 respectively, but aggregated measures n2+n3 and n1+n2+n3 have reliability coefficients higher than 0,74 indicating an acceptable reliability for these indices.

Accuracy

We observe that the reliability of accuracy (accuracy = hits rate – false alarms rate) indices decreases as the load on WM increases, therefore the most reliable accuracy index is the n1 variant. At the same time, in the n1 variant, we observe lower data variability in comparison to other accuracy indices variants. The distribution of accuracy in the n1 variant is strong left-skewed and detailed analysis reveals that 75% of participants have an n1 accuracy score equal or higher than 0,89. Thus, we can conclude that a high reliability coefficient (0,82) in n1 accuracy index variant can result from a ceiling effect. Reliability coefficients for n2 (0,70), n2+n3 (0,78) and n1+n2+n3 (0,83) accuracy variants are acceptable, but the reliability of n3 accuracy (0,61) is not.

Table 2. Descriptive statistics and reliability estimates for indicators calculated from online n-back task

	M	SD	Range	Skewness	Kurtozis	r
Hits rate n1	0,94	0,08	0,34–0,97	-4,33	23,15	0,86
Hits rate n2	0,81	0,15	0,24–0,97	-1,55	2,83	0,69
Hits rate n3	0,65	0,16	0,18–0,97	-0,22	-0,25	0,56
Hits rate n2+n3	0,74	0,14	0,26–0,99	-0,91	1,09	0,75
Hits rate n1+n2+n3	0,81	0,11	0,32–0,99	-1,40	3,01	0,82
False alarms rate n1	0,03	0,03	0,01–0,20	2,48	8,18	0,47
False alarms rate n2	0,06	0,05	0,01–0,32	2,26	8,22	0,62
False alarms rate n3	0,10	0,07	0,01–0,37	1,29	2,07	0,68
False alarms rate n2+n3	0,07	0,05	0,01–0,32	1,71	4,45	0,78
False alarms rate n1+n2+n3	0,06	0,04	0,00–0,28	2,16	7,56	0,75
Accuracy n1	0,91	0,10	0,21–0,96	-3,89	19,85	0,82
Accuracy n2	0,76	0,17	0,10–0,96	-1,39	2,27	0,70
Accuracy n3	0,55	0,19	0,05–0,96	-0,11	-0,57	0,61
Accuracy n2+n3	0,67	0,16	0,12–0,98	-0,71	0,41	0,78
Accuracy n1+n2+n3	0,76	0,13	0,25–0,99	-1,14	1,86	0,83
Response bias n1	0,39	0,19	0,05–0,85	0,45	-0,73	0,36
Response bias n2	0,26	0,15	0,02–0,82	0,94	1,13	0,36
Response bias n3	0,24	0,14	0,01–0,74	0,99	1,13	0,47
Response bias n2+n3	0,23	0,12	0,02–0,66	0,72	1,02	0,43
Response bias n1+n2+n3	0,25	0,12	0,02–0,74	0,83	1,57	0,45
Reaction time n1*	528	115	348–991	1,34	2,54	0,77
Reaction time n2*	618	171	295–1295	1,07	1,50	0,76
Reaction time n3*	704	230	369–1722	1,38	2,83	0,77
Reaction time n2+n3*	645	181	314–1313	1,18	1,63	0,83
Reaction time n1+n2+n3*	578	122	345–1021	0,94	1,08	0,86
Reaction time n1	551	114	374–998	1,22	1,80	0,82
Reaction time n2	661	172	339–1249	0,94	0,81	0,72
Reaction time n3	764	227	382–1699	1,15	2,12	0,76
Reaction time n2+n3	708	174	365–1340	0,94	1,02	0,79
Reaction time n1+n2+n3	646	130	388–1068	0,76	0,41	0,87

M – mean, SD – standard deviation, r – split half reliability calculated as Pearson's correlation corrected with the Spearman Brown formula, n1, n2, n3 – n-back blocks, false alarms rate = (false alarms + 0.5)/(number of distracters + 1.0), hits rate = (hits + 0.5)/(number of targets + 1.0), accuracy = hits rate – false alarms rate, response bias = false alarms rate/(1.0 – accuracy), reaction time – mean or median* for hits only

Response bias

The calculated reliability coefficients for all response bias variants (response bias = false alarms rate /(1,0 – accuracy)) were below 0,5, indicating that we cannot reliably estimate the response bias in any of 5 variants, neither at separate n1, n2, n3 levels nor for aggregated results.

Reaction time

Reaction time (mean or median reaction time for hits) turns out to be the most reliable measure across all analyzed n-back indices; in each variant reaction time reliability coefficients are higher than 0,70, pointing to sufficient reliability. There was no notable difference in reliability of mean and median reaction times. Indices based on median reaction time were slightly more reliable than those based on mean reaction time for n3, n2+n3 and for the n1+n2+n3 variant.

Factorial validity

Due to the fact that the factorial validity of online n-back task has not been tested so far, we decided to investigate it in presented study. In standard n-back task results obtained by participants on different level of "n" (e.g. 1-back, 2-back and 3-back) represent WM functioning under different cognitive load and are separated, but related factors, which together can form a latent variable representing overall WM functioning. We are interested if this theoretical structure could be replicated in online n-back task.

Validity testing was initially based on the analysis of the correlation between two main n-back indicators – accuracy and reaction time; due to their low reliability, we omitted response bias indicators from the validity analysis. Correlations within all 5 variants of accuracy and reaction time are presented in table 3.

In table 3, we can observe that the only index correlating with all calculated reaction times and accuracy indices is the reaction time in n1 blocks. This may suggest that the reaction time index in n1 blocks represents different theoretical constructs than accuracy and other reaction time indices. Therefore, as far as validity is concerned, we proposed a measurement model with 2 separate uncorrelated factors: accuracy – consisting of n1, n2, n3 accuracy indices, and reaction time – consisting of n2, n3 reaction time indices.

To test factorial validity – whether the data collected conformed to a hypothetical two-factor model, we conducted Confirmatory Factor Analysis (CFA). In order to exclude other possible concurrent measurements models, we created 7 different models to evaluate which one fits the data best. The comparison of goodness of fit indices for all models is presented in table 4.

In table 4, we can observe that the assumed two-factor model 1, with an accuracy factor consisting of accuracy indices from n1, n2, n3 blocks and a reaction time factor consisting of reaction time indices from n2 and n3 blocks, shows the best fit to data collected with acceptable values of adjusted goodness of fit index (AGFI), the Tucker-Lewis coefficient (TLI), root mean square error of approximation (RMSA) indices. Four of the eight models, namely 1, 2, 7 and 8, have acceptable TFI and AGFI indices, but only model 1 presents an acceptable upper boundary of a two-tiled 90% confidence interval for RMSA (Schreiber, Nora, Stage, Barlow, & King, 2006). Moreover, for model 1, the CR index is 0,76 for the reaction time factor, and 0,72 for the accuracy factor, indicating an acceptable reliability of the model; AVE is 0,6 for reaction time factor and 0,5 for accuracy factor indicating an acceptable construct validity for both factors. There was no significant correlation between accuracy and reaction time factors in any model.

Table 3. Pearson's correlation between accuracy (Ac) and reaction time (RT) indices

		1	2	3	4	5	6	7	8	9
1	Ac N1	–								
2	Ac N2	**0,46**	–							
3	Ac N3	**0,35**	**0,54**	–						
4	Ac Total	**0,65**	**0,86**	**0,85**	–					
5	Ac N2 N3	**0,46**	**0,86**	**0,89**	**0,97**	–				
6	RT N1	**-0,35**	**-0,22**	**-0,17**	**-0,28**	**-0,22**	–			
7	RT N2	0,03	0,01	0,01	0,02	0,01	**0,50**	–		
8	RT N3	0,07	0,00	0,06	0,05	0,03	**0,23**	**0,54**	–	
9	RT Total	-0,05	-0,04	0,02	-0,02	-0,01	**0,67**	**0,88**	**0,78**	–
10	RT N2 N3	0,06	-0,03	0,04	0,02	0,01	**0,42**	**0,87**	**0,87**	**0,95**

RT – mean reaction time for hits, Ac – accuracy, significant correlation are indicated in bold
p<0,05; 2- tailed

Table 4. Goodness of fit statistics for proposed model (in bold) and 7 concurrent measurement models

Model	X^2	df	p	AGFI	TLI	RMSA	RMSA 90
Ac (N1 N2 N3) </> RT (N2 N3)	**2,0**	**4**	**0,744**	**0,98**	**1,03**	**0,00**	**0,08**
Ac (N1 N2 N3) </> RT* (N2 N3)	0,2	4	0,272	0,96	0,98	0,04	0,13
Ac (N1 N2 N3) </>RT (N1 N2 N3)	35,6	8	0,000	0,84	0,78	0,14	0,19
Ac (N1 N2 N3) </> RT* (N1 N2 N3)	18,0	8	0,022	0,91	0,91	0,09	0,14
Ac (N2 N3) </> RT (N1 N2 N3)	14,4	4	0,006	0,88	0,85	0,12	0,20
Ac (N2 N3) </> RT* (N1 N2 N3)	8,0	4	0,093	0,93	0,93	0,08	0,15
Ac (N2 N3) </> RT (N2 N3)	0,4	1	0,529	0,99	1,03	0,00	0,17
Ac (N2 N3) </> RT* (N2 N3)	0,8	1	0,381	0,98	1,01	0,00	0,19

RT* – median reaction time for hits, RT – mean reaction time for hits, Ac – Accuracy, AGFI – adjusted goodness of fit index, TLI – The Tucker-Lewis coefficient, RMSA – root mean square error of approximation, RMSA 90 – upper boundary of a two-tiled 90% confidence interval for the population RMSEA, </> no significant correlation between latent variables

Discussion

This study addresses the question of whether we can obtain valid and reliable indicators of WM functioning from online n-back task performed on participants' home computers. We conducted the online n-back task among 169 participants and calculated six indicators in 5 variants each (see Table 2 for details).

For hits rate indicators, we can reliably estimate n1, n2+n3 and n1+n2+n3 variants, but the n1 hits rate indicator manifests a ceiling effect represented by low data variability and a high mean score close to a maximum possible result. For false alarms indicators, we can reliably estimate only two aggregated variants, and for accuracy indicators we can reliably estimate all variants, except n3. It might be concluded that, using an online n-back task, we are able to reliably measure overall accuracy – performance success, across 3 n-back blocks: n1, n2, n3, but not accuracy on each n block separately. In other words, we can estimate general performance success over the whole n-back task, representing overall WM functioning, but we cannot reliably compare how participants perform at each of the n levels.

Regarding reaction time, it was the most reliable index from all calculated online n-back task indicators. All variants of two reaction time indicators (mean and median) have acceptable reliability. Therefore, it might be suggested that, using an online n-back task, we are able to reliably estimate reaction time representing performance effort at each n-level, as well as over the whole task.

Our findings are in line with the study of Hockey and Geffen (2004) on the reliability of n-back tasks in laboratory settings (see Table 1). They found that reliability measures were higher in respect to reaction times rather than to accuracy, and concluded that the n-back task is a reliable measure of mental speed. They also found a ceiling effect on the n1 variant. However, in contrast to Hockey and Geffen (2004), in the current study, n3 accuracy variant was not the most, but was the least reliable index from all n-back accuracy indicators. This may be due to a different method of reliability estimating – we used a split half reliability method, and Hockey and Geffen used a one-week test-retest reliability method of estimating. When we consider the work of Jaeggi et al. (2010b) on n-back reliability (see Table 1), where split half correlation was used, we can notice the same pattern of decreasing reliability coefficients with an increase of n level, as in the current study. Jaeggi et al. (2010b) reported mixed results regarding reliability, but reaction time indicators were again more reliable than accuracy measures and a ceiling effect occurred in the n1 variant. As in the current study, but in laboratory settings, Jaeggi et al. (2010b) have shown the lowest reliability to be at n3, out of all of the n levels. This may be explained by an increased error variance appearing with increasing WM load at higher n levels. To sum up, it seems justified to conclude that reliability estimates for accuracy and reaction time in the online n-back task study presented here are generally similar to those obtained in controlled laboratory environments.

Response biases showed the worst reliability of all n-back indicators calculated in this study – across 5 calculated variants of response bias there was no reliable one. In this study, by using the online n-back task, we are unable to reliably estimate participants' performance strategy represented by response bias indicators. This could be due to a relatively low number of false alarms committed by respondents which may produce not enough data points. Probably the experimental procedure (6 targets in one trial x 3 trials x 3 n variants) was too short to capture the respondents' bias tendency.

In order to test factorial validity of aggregated measures we conducted CFA. The best fit to the data collected has been shown to be the model with two uncorrelated factors: accuracy and reaction time. The accuracy factor consisted of results from n1+n2+n3 indicators and reaction time consisted of n2+n3 reaction

time indicators. This two-factor model may be consistent with the proposition of Ragland et al. (2002) which was to consider n-back reaction time indicators as a performance effort measure and accuracy as a performance success measure. These results are also consistent with previous studies (Hockey & Geffen, 2004; Jeaggi et al., 2010b) where no significant correlation between accuracy and reaction time indices was found. Thus, with regard to previous findings and the results of the current CFA, we consider a two-factor measurement model a valid one.

Interestingly, n1 reaction time shows moderate but significant correlation with all of the other reaction time and accuracy indicators. The decision process, whether the previous stimulus was the same as current one, is probably not very demanding for the WM system, thus n1 reaction time may reflect the speed of general mental processes, whereas n2, n3 reaction time indices may additionally reflect the WM system's effort to properly encode the current stimuli under the load. Perhaps n1 reaction time is similar to the inspection time task (Alcorn & Morris, 1996) and may be an indicator of a different theoretical construct than accuracy and reaction time measures (Hockey & Geffen, 2004).

To sum up, by using an online n-back task conducted by participants on home computers, we are able to establish valid and reliable indicators of WM functioning in terms of n-back accuracy and reaction time. Unfortunately, we are unable to reliably establish a response bias indicator and indicators of accuracy for each n block separately. This might be due to overly short experimental blocks in the study.

We initially have shown that we can obtain valid and reliable WM indicators from online studies. However, further studies should concentrate on improving the experimental procedure. We believe that using online n-back tasks may significantly contribute to WM research by decreasing research costs and increasing data sample diversity and size. This study for the first time demonstrates the utility of an online n-back task and may be the first step on a path to popularize experimental WM online research.

Limitations

This study demonstrates that online n-back task conducted on participants' home computers in uncontrolled environment has acceptable psychometric properties similar to n-back task conducted in controlled laboratory settings. However, as an exploratory study, it is not without limitations. Firstly, the research groups consisted of psychology students, who might be more interested in n-back task and highly motivated to perform well, which might have influenced task reliability. Moreover, the recruitment process was not fully online, because students were informed about the study in a classroom by teachers and after this received e-mail message with a link to the online n-back procedure. Thus, recruitment process was partially "stationary", whereas the task performance was online via the Internet. Due to this, the results may not be generalized to people recruited solely via the Internet, without initial physical contact between participants and experimenter. Secondly, we cannot reliably estimate separate accuracy measures for n2, n3 variants and response bias indicators, which might be due to insufficient trial lengths resulting in a small number of data points to analysis. Finally, it is important to note that we analyzed n-back task without lure foils. Our results cannot be generalized to n-back task with lure foils, because, as some researchers point out (Szmalec, Verbruggen, Vandierendonck, & Kemps, 2011), n-back tasks with lures and without lures could be considered to be different tasks. Hence, further research with different types of stimulus and different trial lengths conducted on the general population are needed to replicate these findings. We agree with Gosling and Mason (2015, p. 877) that *Psychological research on the Internet comes with new challenges, but the opportunities far outweigh the costs.*

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs*? Psychological Bulletin, 131*, 30–60.

Alcorn, M., & Morris, G. L. (1996). P300 correlates of inspection time. *Personality and Individual Differences, 20*, 619–627.

Attridge, N., Noonan, D., Eccleston, C., & Keogh, E. (2015). The disruptive effects of pain on n-back task performance in a large general population sample. *Pain*.

Baddeley, A. (2012). Working Memory: Theories, Models, and Controversies. *Annual Review of Psychology, 63*, 1–29.

Baddeley, A. D., & Hitch, G. (1974). Working Memory. *Psychology of Learning and Motivation,* 8, 47–89.

Barchard, K. A., & Williams, J. (2008). Practical advice for conducting ethical online experiments and questionnaires for United States psychologists. *Behavior Research Methods*, *40*(4), 1111–1128. doi:10.3758/BRM.40.4.1111

Byrne, B. (2009). Structural equation modeling with AMOS: Basic concepts, applications, and programming. Uta.Fi. doi:10.4324/9781410600219

Cohen, J. D., Peristein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E. E. (1997). Temporal dynamics of brain activation during a working memory task. *Nature*, *386*, 604–608.

Colom, R., Escorial, Shih, P. C., & Privado, J. (2007). Fluid intelligence, memory span, and temperament difficulties predict academic performance of young adolescent, *Personality and Individual Differences, 42*, 1503–1514.

Conway, A. R., 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*, 769–786.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A metaanalysis. *Psychonomic Bulletin & Review*, *3*, 422–433.

De Clercq, Crombez, Buysse, & Roeyers (2003). A simple and sensitive method to measure timing accuracy. *Behavior Research Methods, Instruments and Computers*, 35, 109–115.

De Dreu, C. K. W., Nijstad, B., Baas, M., Wolsink, I., & Roskes, M. (2012). Working memory benefits creative insight, musical improvisation, and original ideation through maintained task-focused attention. *Personality and Social Psychology Bulletin*, *38*, 656–669.

Domínguez, U., Martín-Rodríguez, J. F., & León-Carrión, J. (2015). Executive n-back tasks for the neuropsychological assessment of working memory. *Behavioural Brain Research, 292*(9), 167–173. doi:10.1016/j.bbr.2015.06.002.

Dougherty, M. R. P., & Hunter, J. (2003). Probability judgment and subadditivity: the role of working memory capacity and constraining retrieval. *Memory & Cognition*, *31*(6), 968–982.

Dougherty, M. R., Hamovitz, T., & Tidwell, J. W. (2015). Reevaluating the effectiveness of n-back training on transfer through the Bayesian lens: Support for the null. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-015-0865-9

Eurostat (2015). Information society statistics. Retrieved from: http://ec.europa.eu/eurostat/web/information-society/data/main-tables [access date: 16.08 2014]

Fornell, C., & Larcker, D. F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error, *Journal of Marketing Research*, *18*(1), 39–50.

Forns, J., Esnaola, M., López-Vicente, M., Suades-González, E., Alvarez-Pedrerol, M., Julvez, J., Grellier, J., Sebastián-Gallés, N., & Sunyer, J. (2014). The n-back test and the attentional network task as measures of child neuropsychological development in epidemiological studies. *Neuropsychology, 28*(4), 519–529. doi:10.1037/neu0000085

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2014). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition, 43*(2), 226–236. doi:10.3758/s13421-014-0461-7

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). Genetic in Origin. *Journal of Experimental Psychology*, *137*(2), 201–225. doi:10.1037/0096-3445.137.2.201

Goozee, R., Reinders, A. A., Handley, R., Marques, T., Taylor, H., O'Daly, O., McQueen, G., Hubbard, K., Mondelli, V., Pariante, C., & Dazzan, P. (2015). Effects of aripiprazole and haloperidol on neural activation during the n-back in healthy individuals: A functional MRI study. *Schizophrenia Research*. doi:10.1016/j.schres.2015.02.023

Gosling, S. D., & Mason, W. (2015). Internet Research in Psychology. *Annual Review of Psychology 66*, 877–902.

Guastello, S. J., Reiter, K., Malon, M., Timm, P., Shircel, A., & Shaline J. (2015). Catastrophe models for cognitive workload and fatigue in N-back tasks. *Nonlinear Dynamics Psychology, and Life Sciences, 19*(2), 173–200.

Hockey, A., & Geffen, G. (2004). The concurrent validity and test-retest reliability of a visuospatial working memory task. *Intelligence*, *32*(6), 591–605. doi:10.1016/j.intell.2004.07.009

Houben, K., & Wiers, R. W. (2008). Measuring implicit alcohol associations via the Internet: validation of Web-based implicit association tests. *Behavior Research Methods*, *40*(4), 1134–1143. doi:10.3758/BRM.40.4.1134

Jacola, L. M., Willard, V. W., Ashford, J. M., Ogg, R. J., Scoggins, M. A., Jones, M. M., Wu, S., & Conklin, H. M. (2014). Clinical utility of the N-back task in functional neuroimaging studies of working memory. *Journal of Clinical and Experimental Neuropsychology*, *36*(8), 875-886. doi:10.1080/13803395.2014.953039

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y. F., Jonides, J., & Perrig, W. J. (2010a). The relationship between n-back performance and matrix reasoning – implications for training and transfer. *Intelligence*, *38*(6), 625–635. doi:10.1016/j.intell.2010.09.001

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010b). The concurrent validity of the N-back task as a working memory measure. *Memory*, *18*(4), 394–412. doi:10.1080/09658211003702171

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (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. doi:10.1037/0278-7393.33.3.615

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

König, C., Bühner, M., & Mürling, G. (2005). Working memory, fluid intelligence, andattention are predictors of multitasking performance, but polychronicity andextraversion are not. *Human Performance*, *18*(3), 243–266.

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological Research Online: Report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet. *American Psychologist, 59*(2), 105–117.

Linck, J., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, *21*, 861–883.

Medow, M. S., Sood, S., Messer, Z., Dzogbeta, S., Terilli, C., & Stewart J. M. (2014). Phenylephrine alteration of cerebral blood flow during orthostasis: effect on n-back performance in chronic fatigue syndrome. *Journal of Applied Physiology*, *117*(10), 1157–1164. doi:10.1152/japplphysiol.00527.2014

Miller, K. M., Price, C. C., Okun, M. S., Montijo, H., & Bowers, D. (2009). Is the N-back task a valid neuropsychological measure for assessing working memory? *Archives of Clinical Neuropsychology*, *24*(7), 711–717. doi:10.1093/arclin/acp063

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). E-Research: Ethics, Security, Design, and Control in Psychological Research on the Internet. *Journal of Social Issues*, *58*(1), 161–176.

Oberauer, K., Schulze, R., Wilhelm, O., & Süss, 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; author reply 72–75.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittman, W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence, 31*, 167–193.

Pérez, A. I., Paolieri, D., Macizo, P., & Bajo, T. (2014). The role of working memory in inferential sentence comprehension. *Cognitive Processing, 15*(3), 405–413.

Philip, N. S., Sweet, L. H., Tyrka, A. R., Carpenter, S. L., Albright, S. E., Price, L. H., & Carpenter, L. L. (2015). Exposure to childhood trauma is associated with altered n-back activation and performance in healthy adults: implications for a commonly used working memory task. *Brain Imaging Behavior*.

Ragland, J. D., Turetsky, B. I., Gur, R. C., Gunning-Dixon, F., Turner, T., Schroeder, L., Chan, R., & Gur, R. E. (2002). Working memory for complex figures: an fMRI comparison of letter and fractal n-back tasks. *Neuropsychology*, *16*(3), 370–379. doi:10.1037/0894-4105.16.3.370

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*(2), 173–184.

Raz, S., Bar-Haim, Y., Sadeh, A., & Dan, O. (2012). Reliability and Validity of the Online Continuous Performance Test Among Young Adults. *Assessment*, *21*(1), 108-118. doi:10.1177/1073191112443409

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

Salminen, T., Frensch, P., Strobach, T., & Schubert, T. (2015). Age-specific differences of dual n-back training. *Neuropsychology, development, and cognition. Section B, Aging, neuropsychology and cognition*, *13*, 1-22. doi:10.1080/13825585.2015.1031723

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. doi:10.1037/0096-3445.132.4.566

Sánchez-Fernández, J., Muñoz-Leiva, F., & Montoro-Ríos, F. J. (2012). Improving retention rate and response quality in Web-based surveys. *Computers in Human Behavior*, *28*(2), 507–514. doi:10.1016/j.chb.2011.10.023

Schmiedek, F., Hildebrandt, A., Lövdén, M., Lindenberger, U., & Wilhelm, O. (2009). Complex span versus updating tasks of working memory: the gap is not that deep. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(4), 1089–1096. doi:10.1037/a0015730

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2014). A task is a task is a task: putting complex span, n-back, and other working memory indicators in psychometric context. *Frontiers in Psychology*, *5*, 1–8. doi:10.3389/fpsyg.2014.01475

Shelton, J. T., Elliott, E. M., Hill, B., Calamia, M. R., & Gouvier, W. (2009). A comparison of laboratory and clinical working memory tests and their prediction of fluid intelligence. *Intelligence*, *37*(3), 283–293. doi:10.1016/j.intell.2008.11.005

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. doi:10.1037/0096-3445.117.1.34

Sörqvist, P. (2010). The role of working memory capacity in auditory distraction: a review. *Noise & Health*, *12,* 217–224.

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, 137–151.

Unsworth, N., Brewer, G., & Spillers, G. J. (2012). Working memory capacity and retrieval from long-term memory: the role of controlled search. *Memory & Cognition, 41*(2), 242–254.

Van de Weijer-Bergsma, E., Kroesbergen, E. H., Prast, E. J., & Van Luit, J. E. H. (2014). Validity and reliability of an online visual-spatial working memory task for self-reliant administration in school-aged children. *Behavior Research Methods*, *47*(3), 708–719. doi:10.3758/s13428-014-0469-8

Van Leeuwen, M., van den Berg, S. M., Hoekstra, R. A., & Boomsma, D. I. (2007). Endophenotypes for intelligence in children and adolescents. *Intelligence*, *35*(4), 369–380. doi:10.1016/j.intell.2006.09.008

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*, 1–22. doi:10.3389/fpsyg.2013.0043